Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 Sonbahar 2017 Cilt: 8- Sayı: 3

Autumn 2017 Volume: 8- Issue: 3

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluluğu ilgili yazarlara aittir.

İletişim
e-posta: epod@epod-online.org
Web: http://epod-online.org

Journal of Measurement and Evaluation in Education and Psychology (EPOD) is a national refereed journal that is published four times a year. The responsibility lies with the authors of papers.

Contact
e-mail: epod@epod-online.org
Web: http://epod-online.o

Dizinleme / Abstracting & Indexing

DOAJ (Directory of Open Access Journals), TÜBİTAK Ulakbim Sosyal ve Beşeri Bilimler Veri Tabanı, Tei (Türk Eğitim İndeksi)

Hakem Kurulu / Referee Board

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Eylül 2017, Sayı: 8-3)

Journal of Measurement and Evaluation in Education and Psychology (September 2017, Issue: 8-3)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);246-265

8. Sınıf Matematik Akademik Başarısını Yordayan Faktörler-TIMSS 2015

Factors Predicting Mathematics Achievement of 8th Graders in TIMSS 2015

Mehmet Hayri SARI * Serkan ARIKAN ** Hülya YILDIZLI ***

Öz

Bu çalışmada TIMSS 2015 uygulamasında ele alınan öğrenci, öğretmen ve okul faktörlerinin Türkiye'deki sekizinci sınıf öğrencilerinin matematik başarıları ile nasıl bir ilişki içinde olduğunun araştırılması amaçlanmıştır. İlişkisel tarama modeline göre tasarlanan araştırmanın çalışma grubunu, Uluslararası Matematik ve Fen Eğilimleri Araştırması 2015 uygulamasına katılan 6079 öğrenci ve 220 öğretmen oluşturmuştur. Araştırmanın veri toplama araçları, TIMSS 2015 uygulamasında elde edilen öğrenci ve öğretmen anketleri ile matematik başarı testinden oluşmaktadır. Verilerin analizinde, matematik başarı puanları bağımlı değişken; öğrenci ve öğretmen özellikleri değişkenleri bağımsız değişken olacak şekilde hiyerarşik regresyon analizi yöntemi kullanılmıştır. Araştırmadan elde edilen bulgulara göre, öğrenci ile ilgili değişkenlerin TIMSS 2015'teki matematik başarısındaki farklılıkların %34'ünü açıkladığı görülmüştür. Duyuşsal alan boyutunda yer alan öz-yeterlik inancı 8.sınıf öğrencilerinin TIMSS 2015'te matematik başarılarını yordamada en önemli değişkendir. Öz-yeterlik inancından sonra matematik başarısını yordamada diğer bir önemli değişken öğrencilerin evde sahip oldukları eğitimsel kaynakları olmuştur. Duyuşsal alan içerisinde yer alan tutum ile matematik başarısı arasında negatif yönde bir ilişki bulunmuştur. Matematiğe verilen önem ile öğrencilerin matematik başarıları arasında ise anlamlı bir ilişki çıkmamıştır. Araştırmada okul algısı boyutu altında ele alınan zorbalık, okula aidiyet ve öğretim etkinlikleri ile matematik başarısı arasında anlamlı bir ilişki olup bu ilişki göreli olarak matematik başarısını yordamada daha az öneme sahiptir. Öğrenmede çevre faktörü altında ele alınan okulda başarıya verilen önem, güvenli ve düzenli okul ortamı, okul koşulları, iş tatmini, öğretmenin karşılaştığı sorunlar ve öğrencilerden kaynaklı sorunların hepsi bir araya geldiğinde okullardaki 8. sınıflar arası başarı farkının %29'unu açıklamaktadır. Öğrenmede çevre faktöründe yer alan okulda başarıya verilen önem ve öğrencilerden kaynaklı sorunlar öğrencilerin matematik başarısını yordamada iki önemli değişken olmuştur.

Anahtar Kelimeler: TIMSS, matematik başarısı, duyuşsal alan, çevresel faktör, öğretmen

Abstract

In the study, it is aimed to investigate the student, teacher and school factors predicting mathematics achievement of Turkish 8th grade students in TIMSS 2015. The group of the study consists of 6079 students and 220 teachers who attended TIMSS from Turkey. The data of the study was obtained from student and teacher questionnaires and mathematics cognitive test scores. In the data analysis, multilevel regression analysis was used in which dependent variables were plausible mathematics scores and independent variables were student, teacher and school scale scores. According to results, 34% percent of student-level variance was explained by student-level variables. It was found that self-confidence level of students was the most important predictor of mathematics achievement among student-level variables. Additionally, educational resources at home variable was also among the important predictors of mathematics achievement. Teacher and school factors explained 29% of between school variance. Among these variables, school emphasis on academic success and teaching limited by student needs were two significant variables that could predict mathematics achievement of students.

* Yrd. Doç. Dr., Nevşehir Hacı Bektaş Veli Üniversitesi, Eğitim Fakültesi, Nevşehir-Türkiye, mhsari@nevsehir.edu.tr ,ORCID ID: orcid.org/0000-0002-7159-2635

** Yrd. Doç. Dr., Muğla Sıtkı Koçman Üniversitesi, Eğitim Fakültesi, Muğla-Türkiye, serkanarikan@mu.edu.tr ,ORCID ID: orcid.org/0000-0001-9610-5496

*** Yrd. Doç. Dr., Nevşehir Hacı Bektaş Veli Üniversitesi, Eğitim Fakültesi, Nevşehir-Türkiye, hulyayildizli@nevsehir.edu.tr, ORCID ID: orcid.org/0000-0003-4450-2128

Geliş Tarihi: 04.03.2017
DOI: 10.21031/epod.303689
Kabul Tarihi: 09.06.2017

Keywords: TIMSS, mathematics achievement, affective domain, environmental factor, teacher

GİRİŞ

Günümüze kadar bireyin öğrenmesiyle ilgili çeşitli kuramlar geliştirilmiş ve bu kuramlar çerçevesinde öğrenmenin nasıl olduğuna yönelik farklı bakış açıları geliştirilerek farklı tanımlamalar yapılmıştır. Bu tanımlamalar bağlamında öğrenmede etkili olduğu düşünülen faktörler, kuramların bakış açılarına göre ve bireyin öğrenme sürecindeki etkinliğine göre değişmiştir. Öğrenmeyi açıklayan önemli kuramlardan biri olan bilişsel kuramı destekleyenler, çevresel değişkenleri ve bu değişkenlerin bireye etkisini göz önünde bulundurmadan sadece bilgiyi kodlama, işleme, kaydetme ve geri getirme gibi zihinsel süreçlerle ilgilenmiş (Schunk, 2008), üst-biliş ve sosyal-biliş alanındaki araştırmacılar ise öğrenmede bireysel farklılıklar ve farklılıkların öğrenme ile ilişkisi konularına yeni bakış açıları getirmeye çalışmışlardır (Zimmerman, 2002). Bu gelişmelerle birlikte öğrenmede sadece bilişsel faktörlerin ve süreçlerin değil, aynı zamanda çevresel faktörlerin (Bandura, 1986; Wood ve Bandura, 1989) ve duyuşsal faktörlerin var olduğu (Bloom, 2012), bu faktörler arasında etkileşimin önemli olduğu ve bu etkileşimler sonucunda öğrenmelerin düzenlenmekte ve açıklanmakta olduğu ifade edilmiştir (Bandura, 1986; Wood ve Bandura, 1989). Bireyin nasıl öğrendiğinin açıklanmasında öğrencilerin duyuşsal özellikleri, kendisi ile ilgili faktörler (cinsiyet, ailenin yapısı ve ilişkileri, bireysel yetenekler, büyüme/gelişme özellikleri, ihtiyaçlar), öğrenme ortamı (çevresel faktörler) ve bu öğrenme ortamı içindeki değişkenler (aile, okul, sınıf, öğretmen, arkadaş vb.) de önem kazanmıştır.

Öğrenmede önemli olan bu değişkenlerin matematik, fen bilimleri ve okuma gibi alanlar ile nasıl bir ilişki içinde oldukları uluslararası düzeyde yapılan sınavlarla ölçülmekte ve ülkelerin bu alanlara yönelik eğitim durumları istatistiksel olarak ortaya çıkarılmaktadır. Türkiye'nin de katıldığı, öğretmen ve öğrenci nitelikleri üzerine değerlendirmelerin yapıldığı TIMSS (Trends in International Mathematics and Science Study; Uluslararası Matematik ve Fen Eğilimleri Araştırması) eğitim sistemimizin iyi bir durumda olmadığını ve sistemde önemli adımlar atılması gerektiğini ortaya çıkarmaktadır. TIMSS 2015 raporuna göre Türkiye; matematik alanında 4. sınıflarda 36. sırada, 8. sınıflarda ise 24. sırada yer almaktadır (Yücel ve Karadağ, 2016). Bu çalışmada da TIMSS 2015 çalışma grubunda yer alan 8. sınıf öğrencilerinin duyuşsal ve çevresel faktörlerinin matematik başarıları ile nasıl bir ilişki içinde olduğu araştırılmaktadır. Bu araştırmanın sonuçları göz önüne alındığında matematik başarısı ile anlamlı ilişki içerisinde olan değişkenleri belirlemek ve bunları anlamlandırmak başta ulusal liderlere, eğitim politikacılarına ve eğitimcilere yön göstermesi açısından önemli görülmektedir (Yavuz, Demirtaşlı, Yalçın ve İlgün Dibek, 2017).

Öğrenmede Öğrenci

Duyuşsal alan

Öğrenmede anahtar bir rolü olan duyuşsal alan akademik olarak bireylerin başarılı olabilmeleri için gerekli eğilimlerden biri olarak karşımıza çıkmaktadır. Bireylerin öğrenme sürecinde sorumluluk sahibi olma, işbirliği yapabilme, engeller karşısında sebat etme, değer verme, tutum, kendine güven vb. özelliklere de ihtiyacı vardır (Mcmillian, 2015). Bu özellikler öğrencilerin düşünme, anlama, harekete geçme durumlarında önemli görülmekte ve öğrencilerin öğrenmeleri arasında yaşanan farklılıkların kaynağının yaklaşık dörtte birini duyuşsal alan oluşturduğu ifade edilmektedir (Bloom, 2012).

Genel eğitim sisteminde duyuşsal alana yer verilmesine rağmen mevcut sınav sistemleri bilişsel yeterliklerin ön plana çıkarılmasına ve öğretmenlerin öğretim programlarında bilişsel alan becerilerine daha çok odaklanmalarına sebep olmuştur. Bu durum eğitim ortamlarında duyuşsal alandaki gelişimin büyük ölçüde rastlantılara bırakıldığını vurgulamaktadır (Şimşek, 2009). Hâlbuki duyuşsal alan ile başarı arasındaki ilişkiyi araştıran çalışmalardan elde edilen bulgular, duyuşsal alanın öğrencilerin başarılarını belirlemede ve yordamada önemli bir rolü olduğunu göstermektedir (Ferla, Valcke ve Cai, 2009; Leder ve Forgasz, 2006; Pajares ve Miller, 1997). Özellikle matematik öğrenme ve öğretme sürecinde tutum ve inançlarla ilgili matematiksel eğilimlerin öğrencilerin günlük hayatında bu bilgiyi bilinçli bir şekilde kullanmasında ve bu bilgiyi kullanmada istekli olmasında alan bilgisi kadar büyük öneme sahip olduğu belirtilmektedir (Wilkins ve Ma, 2003). Bu ilişkilerden yola çıkarak, yapılan uluslararası sınavlarda ülkemizin düşük başarısının nedenleri arasında öğretim yaşantılarının yetersizliğinin yanı sıra öğrenenlerin matematik dersine yönelik psikolojik tepkileri de gösterilmektedir (Tuncer ve Yılmaz, 2016).

Başarı ile ilişki gösteren duyuşsal alana ait özelliklerden ilki tutumdur. Tutum, herhangi bir objeye, kişiye veya kuruma karşı olumlu ya da olumsuz bir tepkide bulunma, olumsuz tepkiyle birlikte o objeye karşı ilgisiz kalma ve objeye karşı lehte ya da aleyhte gerçekleşen duygusal eğilim olarak tanımlanmaktadır (Papanastasiou, 2002; Turgut ve Baykul, 2012). Bu araştırmada tutum değişkeninin ele alınmasının iki gerekçesi vardır: Birincisi, başarıda gözlenen toplam değişkenliğin %12 ile %20 arasındaki bir kısmı derse yönelik tutumdaki farklarla açıklanabilmektedir (Bloom, 2012). Diğer bir ifadeyle, öğrencinin bir alana yönelik tutumu istenen düzeyde ise belirlenen hedef davranışları kazanması daha kolay olacak ve söz konusu matematik dersi olunca matematiğe yönelik olumlu tutum geliştirme öğrencinin matematik öğrenmeye hazır hale gelmesini sağlayacaktır (Demir ve Kılıç, 2010). İkincisi ise, TIMSS ve PISA sınavlarında Türkiye'deki öğrencilerin matematik dersine yönelik tutumlarının matematik başarısı ile ilişkisinin farklı sonuçlar ortaya koymasıdır. Örneğin; Doğan ve Barış'ın (2010) TIMSS 2007 verilerini kullandıkları çalışmalarında öğrencilerin matematik dersine yönelik tutumları ile matematik başarıları arasında negatif yönde ve anlamlı bir ilişki tespit ederken; TIMSS 1999 sınavında tutum değişkeninin öğrencilerin matematik başarıları üzerinde anlamlı bir etkisi bulunmadığını belirlemişlerdir. Aynı şekilde TIMSS 2011 sınavında tutum değişkeni ile uygulamaya katılan öğrencilerin matematik başarıları arasında negatif bir ilişki olduğu ortaya konulmuştur (Yavuz vd., 2017). Benzer sonuçlar ulusal verilerle yapılan diğer çalışmalarla da ortaya konulmuştur (Ekizoğlu ve Tezer, 2007; Peker ve Mirasyedioğlu, 2003; Uygun ve Işık Tertemiz, 2014).

Duyuşsal alan değişkenlerinden bir diğeri öz-yeterlik inancıdır. Öz-yeterlik, "bireyin belli bir performansı gösterebilmesinde olası durumlar ile başa çıkabilmek için gerekli olan eylemleri ne kadar iyi yapabildiğine ilişkin yargılarıdır" (Bandura, 1986; s. 391). Başka bir tanıma göre öz-yeterlik, bireyin gerçekleştirmesi gereken performans ile kendi kapasitesini karşılaştırıp, duruma göre harekete geçmesi ve bireyin karşılaşmış olduğu güçlüklerde nasıl başarılı olabileceğine ilişkin kendisi hakkındaki inancıdır (Bayrakçı, 2007). Bu nedenle bireylerin bildikleri şeylerin ve sahip oldukları becerilerin uygulanması öz-yeterlikten etkilenmektedir (Wilson ve Narayan, 2016). Öz-yeterlik, öğrenme sürecinde araştırılması gereken bir değişken olarak görülmekte ve matematik başarısında bir öngörücü, arabulucu ve bir performans sonucu olarak rol oynadığı belirtilmektedir (Wilson ve Narayan, 2016). Öğrencilerin kendilerine yönelik yeterlik algılarının matematik başarısını etkilediği (Usta, 2014) ve matematiğin önemli becerilerinden biri olan problem çözme becerilerinin matematik öz-yeterliğiyle anlamlı ve yüksek bir ilişkisi olduğu belirtilmektedir (Çelik, 2012; Pajares ve Miller, 1997). Yapılan araştırmalarda (Ayan, 2014; Chen, 2003; Organisation for Economic Co-Operation and Development [OECD], 2004; Özgen ve Bindak, 2011; Özüdoğru, 2013), öz-yeterlik, matematik akademik başarısının ve matematik okuryazarlığının önemli bir yordayıcısı olarak görülmektedir.

Duyuşsal alan özelliklerinden bir diğeri ise öğrenme değeridir. Matematik öğrenmeye değer verme, öğrencilerin matematiğin önemine ve matematiğin yaşamlarının çeşitli dönemlerindeki faydasına yönelik öğrencilerin geliştirmiş oldukları tutumunu ifade etmektedir (Wigfield ve Eccles, 1992). Değer vermede öğrenmenin ve bu süreçte bireyin üzerine düşen görevlerin önemine ilişkin bireyin algısı, kişisel ilgi, öğrenmenin bireyin kendisine fayda getirmesi ve bireyin gelecekteki hedeflerine yönelik öğrenme değerinin faydasına ilişkin algısı önem arz etmektedir (Wigfield ve Eccless, 1992). Buna karşın, 2007 ve 2011 yıllarında yapılan TIMSS sınavlarında Türk öğrencilerinin matematiğe verdiği değer değişkeninin her iki dönemdeki sınavda öğrencilerin matematik başarı puanları ile anlamlı bir ilişkiye sahip olmadığı bulunmuştur (Arıkan, van de Vijver ve Yağmur, 2016; Yavuz vd., 2017). Hâlbuki bireyin matematiksel öğrenmeleri hakkında olumlu değerlendirmeler yaparak bu görevin kendisi için önemli olduğunu düşünmesi öğrenmede gerekli bilişsel, üstbilişsel ve duyuşsal stratejileri daha etkin ve bilinçli kullanmasını sağlayacak, bu durum matematik başarısını etkileyebilecektir. Bu nedenle 2007 ve 2011 yıllarından sonra 2015 yılında yapılan TIMSS sınavında öğrencilerin matematiğe verdikleri değer ile matematik başarıları arasındaki ilişkinin araştırılmasının önemli olduğu düşünülmektedir.

Okul algısı

Okul boyutunda yer alan öğrenci-öğrenci ilişkileri bağlamında alınabilecek değişkenlerden biri zorbalıktır. Zorbalık (Bullying), öğrencilerin gereksiz, yıkıcı ve güce dayalı davranışlarıdır. Zorbalığın ad takma, başkalarına vurma veya tehdit etme ve yanlış söylentiler yayma gibi çok sayıda biçimi bulunmaktadır. Öğrenciler arasında zorbalık, sadece akademik performanslarını düşürmekle kalmaz aynı zamanda zihinsel sağlık sorunlarına ve fiziksel yaralanmaya da neden olabilir (Jan ve Husain, 2015). Okul iklimi içerisinde gösterilen zorbalık, matematik başarısı üzerinde önemli bir etkiye sahiptir (Buluç, 2014; Lee Van Horn, 2003; Mohammadpour, 2012; Yavuz, vd., 2017). Şiddet ve zorbalık olayları okul iklimini olumsuz yönde etkilemekte, bu olaylar diğer ülkelere oranla Türkiye'de daha fazla görülmekte ve bu durum öğrencilerin matematik başarısına olumsuz yansımaktadır (Buluç, 2014). Araştırmalar yüksek zorbalığa maruz kalma düzeyine sahip olan okullara devam eden öğrencilerin okuldaki performanslarının daha düşük olabileceğini göstermektedir (Strøm, Thoresen, Wentzel-Larsen ve Dyb, 2013). Milli Eğitim Bakanlığının [MEB] 2016 yılında yayınlamış olduğu bir raporda araştırmaya katılan 277.000 öğretmenin %21'i öğrencilerin akran baskısı, zorbalık, şiddet vb. nedenlerden ötürü devamsızlık yapabileceklerini belirtmektedir (MEB, 2016). 2007 ve 2011 yıllarında yapılan TIMSS sınavlarında okullarında zorbalığa maruz kalmayan öğrencilerin maruz kalan öğrencilere göre matematik başarıları daha yüksek çıkmıştır. (Ölçüoğlu ve Çetin, 2016; Yavuz, vd., 2017). Bu nedenle TIMSS 2015 sınavında öğrencilerin zorbalığa maruz kalma dereceleri ile matematik başarıları arasında bir ilişkinin devam edip etmediğini incelemek gerekmektedir.

Araştırmada ele alınan diğer bir çevresel özellik ise okula aidiyettir. Okula aidiyet bireyin akranlar, öğretmenler ve okullarla bir kabul, katılım ve bağlantı duygusu yani öğrencinin okulundaki diğer öğrenciler tarafından ne ölçüde kabullenildiğine, saygı duyulduğuna, dâhil edildiğine ve desteklendiğine yönelik kişisel duygu durumu olarak tanımlanmıştır (Goodenow ve Grady, 1993). Olumlu okul ortamları ve okul aidiyeti, öğrenciler arasında pozitif yönde akademik, sosyal ve psikolojik sonuçlar ile ilişkilendirilmektedir (McMahon, Parnes, Keys ve Viola, 2008). Öğrencilerin okuluna ya da bulunduğu sınıf ortamına karşı geliştirmiş oldukları aidiyet duygusu akademik başarıyı etkileyen önemli faktörlerden biridir (Anderson, 2010; Duru ve Balkıs, 2015). Başka bir ifadeyle; öğrencilerin aidiyet algıları ile okul ortamındaki başarı arasında ilişki olduğu ifade edilmektedir (Anderson, 2010; Nichols, 2008). Matematik dersine yönelik olarak daha önce yapılan çalışmalarda öğrencilerin okula yönelik geliştirmiş oldukları aidiyet duygusunun matematik başarısını anlamlı düzeyde yordadığı bulunmuştur (Phan, 2013; Akt. Duru ve Balkıs, 2015).

TIMSS 2015 değişkenlerinden olan ve matematik başarısı ile ilişkisi olup olmadığı araştırılan bir diğer faktör ise öğrencilerin matematik öğretim etkinlikleri hakkındaki görüşleridir. Öğrenme ortamında öğretmenin mesleki ve kişisel yeterlikleri öğrenme kalitesini etkilemektedir. Öğretmen ve öğrenci sürekli etkileşim halinde olduğundan birinde yaşanan olumsuzluk öğrenme sürecinin sağlıklı bir şekilde gitmesine engel olmaktadır. Öğrencinin öğrenme sürecinde verimli olması sadece kendi çabasıyla değil, öğretmenin davranış biçimi, bilgi ve becerisinden de etkilenebilmektedir (Aydın, 1993). Matematik öğretiminde yaşam boyu öğrenmeyi, problem çözme yeteneğini geliştirmeyi ve üst düzey düşünme becerilerini kazandırmayı amaçlayan yeni yönelimler ve yaklaşımlar sayesinde öğrenci aktivitesi ön plana çıkmıştır. Öğrenci aktivitelerinde matematiksel kavramlar kazandırılmaktan ziyade öğrenciyi matematik öğrenmeye istekli ve ilgili hale getirmek amaçlanmalı ve bu süreçte öğretmen-öğrenci arasında etkili iletişim sağlanarak öğrenmenin anlamlandırılması sağlanmalıdır. Matematik dersinde bu yaklaşımların uygulanması ile öğrenciler daha kalıcı öğrenebilmekte ve matematiğe yönelik ilgileri ve diğer duyuşsal alanları olumlu anlamda etkilenebilmektedir. TIMSS 1999 ve TIMSS 2007 karşılaştırılması yapılan bir araştırmada öğrencilerin matematik dersini günlük yaşamla ilişkilendirme yüzdelerinin arttığı görülmüştür. Bu durum yeni öğretim yaklaşımlarının sınıflarda daha fazla uygulandığını gösterebilir (Bilican, Demirtaşlı ve Kilmen, 2011). Bu nedenle şu an uygulanmakta olan matematik öğretim programlarının ve bu bağlamda sınıf içi yeni yönelim ve yaklaşımların matematik akademik başarısına önemli bir katkısının olup olmadığının araştırılmasının önemli olduğu düşünülmektedir.

Öğrencilerin evdeki kaynakları

Okulun öğrenme çevresi olarak etkin bir rolde yer alabilmesi için kendi dışındaki çevreyle iç içe olduğunun farkında olması ve belirlediği hedeflerine ulaşmada bulunduğu çevreyle olan ilişkilerini geliştirmesi gerekmektedir (Keçeli Kaysılı, 2008). Okulun sahip olduğu çevrelerden biri de öğrencilerin aileleridir. Aile öğrencinin okulda öğrendikleri ile okul dışındaki fırsatları arasında bağlantı kurmada önemli hale gelmektedir (Danielson, 2002). Bu fırsatlar altında sayılabilecek değişkenler; evde bulunan kitap sayısı, bilgisayar, anne-babanın sosyo-ekonomik durumu, ailenin eğitime katılma düzeyi akademik başarıda önemli görülmektedir. Daha önceki yıllarda yapılan TIMSS ve PISA sınavlarının sonuçlarına bakıldığında akademik başarı ile destekleyici ev ortamı arasında güçlü ve pozitif bir ilişkinin varlığı (Bayar ve Bayar, 2013), matematik başarısı ile güçlü bir ilişkide olduğu (Brese ve Mirazchiyski, 2010, Ölçüoğlu ve Çetin, 2016; Özer ve Anıl, 2011; Yılmaz ve Bindak, 2016) görülmektedir. Türkiye'de son yıllarda gelir düzeyi yüksek olan aileler çocuklarını özel okullara göndermekte ya da okul dışı öğretim faaliyetlerini artırmakta, bununla birlikte evdeki kaynakları öğretime daha fazla hizmet eder hale getirmeye çalışmaktadır. Bu sebeple ailelerin gelir düzeyleri arasındaki farklardan doğan eşitsizlikler daha da artmakta, bu durum öğrencilerin evde sahip olduğu kaynakların her geçen gün daha önemli hale gelmesine yol açmaktadır. TIMSS 2015 matematik başarısını yordamada evdeki kaynakların önemli olup olmadığının incelenmesi önemli görülmektedir.

Öğrenmede Çevre (Okul-Öğretmen)

Eğitimde çevre, öğrenmenin daha verimli olmasında ve dolayısıyla öğretimin etkili bir şekilde gerçekleşmesinde en önemli öğelerden biridir (Büyükkaragöz ve Çivi, 1999). Öğrenmede çevre değişkeni öğrencilerin matematik öğrenmelerinde hem akademik başarıda hem de matematiğe yönelik olumlu duygusal deneyimler kazanmasında etkili olmaktadır (Frenzel, Pekrun ve Goetz, 2007). Eğitim sistemi içinde belirlenen amaçların somut birtakım faaliyetlere dönüştüğü alanlar okullardır (Ekinci, 2014). Eğitimin önemli bir boyutunu oluşturan okulun güvenli ve düzenli olması, eğitim-öğretim faaliyetlerinin korkudan, şiddetten ve kaygıdan uzak hoş bir ortam olmasıdır (Çalık, Kurt ve Çalık, 2011). Böyle ortamlar, her öğrenciye özen ve kabul duygusunun ve olumlu bir atmosferin hâkim olduğu eğitimsel bir iklim sağlar (Çalık vd., 2011; Lezotte, 1993). Türkiye, TIMSS 2011 raporunda güvenli ve düzenli okul ortamının sağlanması anketine verilen puan ortalamasında 50 ülke içinden 43. sırada yer almıştır. Öğretmenlerin değerlendirmelerine bakıldığında Türkiye örneklem grubunda yer alan öğrencilerin öğrenim gördükleri okulları çok fazla güvenli ve düzenli bulmadıklarını ifade etmişlerdir (Buluç, 2014). Benzer şekilde başka bir araştırmada velilerin büyük çoğunluğunun (%64) okulun güvenliği konusunda endişe duydukları ortaya konulmuştur (İksara, 2013). Hâlbuki okulda ortamında sağlanan güven ve huzur, günümüz eğitim yönetimi paradigmasında önemli bir kavram olarak karşımıza çıkmakta ve öğrencinin okuldaki güvenilir bir ortamdan dolayı yaşadığı rahatlık, onun okul yaşamına yönelik motivasyonunu da önemli düzeyde etkileyebilmektedir (Yaman, Eroğlu, Bayraktar ve Çolak, 2010). Bununla birlikte okulun sahip olduğu koşulların (temizlik, teknolojik donanım, öğretim materyallerinin sayısı vb.) öğrenci başarısı üzerinde önemli bir etkiye sahip olduğu belirtilmektedir (American Federation of Teachers, 2006). Yapılan birçok araştırmada güvenli ve düzenli okul ortamının, okulda başarıya verilen önemin (Buluç, 2014; Ölçüoğlu, 2015) ve okulun sahip olduğu koşulların (Baker ve Bernstein, 2012; Clark, 2002; Karasolak ve Sarı, 2011) öğretmen-öğrenci davranışları ile akademik başarı üzerinde fark edilebilir bir etkisinin olduğu ortaya çıkmıştır.

Okul ortamıyla birlikte sınıf ortamı ve bu bağlamda öğrenci profili öğretmenin sınıf içi uygulamaları üzerinde etkili olabilmektedir. Örneğin; öğrencinin zihinsel ve fiziksel yeterlik/yetersizliği, öğrenmeye istekli olma/olmama durumu, istenmeyen davranışlar (dersi bölme, sorumluluk almaktan kaçma) gibi değişkenler öğretim sürecine yansımaktadır. Türkiye'de yapılan araştırmalar incelendiğinde öğretmenlerin çok fazla sorunlara sahip olduğu görülmektedir (Demir ve Arı, 2013; Ekinci, 2014; Karacaoğlu ve Kaçar, 2010). Bu sorunlar programları uygulamada sürenin yetmediği, ders araç-gereçlerin yetersizliği, çalışma saatlerinin fazlalığı, öğrenci sayısının fazlalığı ve fiziki imkânlara dayanan zorluklar, okul yönetimi ile ilgili sorunlar ve bürokrasi işlerinin fazlalığıdır (Demir ve Arı, 2013; Karacaoğlu ve Kaçar, 2010). Bununla birlikte öğrencinin sahip olduğu ailede kültür, sosyo-ekonomik durum, çocuğa karşı davranış ve tutumlar, aile üyelerinin çocuklarına verdiği değer gibi sorunlar öğretmenlerce ifade edilmekte ve bu sorunlar öğrencilerin okuldaki başarılarını büyük ölçüde etkilemektedir (Ekinci, 2014; Engin, Özen ve Bayoğlu, 2009).

Okul ve sınıf içindeki olumlu ya da olumsuz yaşantılar öğretmenlerin mesleklerine yönelik farklı duygu ve düşünce oluşturmalarına sebep olabilmektedir. Bu duygu ve düşünceler öğretmenin mesleği ile ilgili yaşantısına yönelik, memnuniyet verme-vermeme veya olumlu bir duygu ile sonuçlanma-sonuçlanmama gibi yargılara varmasına sebep olmaktadır. Burada mesleğe yönelik iş tatmini kavramı önümüze çıkmaktadır. "İş tatmini, bireylerin iş tecrübeleri sonucunda elde ettiği olumlu ruh hali" (Erdoğan, 1996; s. 231) olup öğretmenin mesleğini icra ederken yaptığı işe karşı hissettikleri olarak tanımlanabilir. Öğretmenin iş tatmini, mesleğe yönelik memnuniyeti öğretim uygulamalarına yansımakta ve bu durum da öğrenci başarısı ile ilişkili hale gelmektedir. Çünkü mesleklerinden memnun olan öğretmenler üzerlerine düşen görev ve sorumlulukları büyük ihtimalle etkili bir şekilde yerine getirebileceklerdir. PISA 2012 ve TIMSS 2011 sonuçlarına göre öğrencilerin matematik alanlarında akademik başarılarının yüksek olduğu ülkelerde öğretmen motivasyonu diğer ülkelere göre daha yüksek olduğu görülmektedir (Abazaoğlu ve Aztekin, 2015). Türkiye'deki öğretmenlerin iş tatmini düzeylerinin öğrencilerinin başarılarıyla ilişkili olup olmadığının belirlenmesi mesleğin geleceğine yönelik atılacak adımlarda yön göstermesi bakımından önem arz etmektedir.

Yukarıda bahsedilen öğrenci, öğretmen ve okul özellikleri dikkate alındığında bireyin başarısını tek bir değişkenin yordamadığı ve birden fazla değişkenin başarının açıklanmasında önemli rolleri olduğu görülmektedir. Özellikle TIMSS gibi uluslararası kapsamda ve geniş çalışma grupları üzerinde yapılan sınavlar ülkelerin eğitim sistemlerini değerlendirmede, öğrencilerin fen, matematik, okuma gibi alanlardaki başarılarını açıklamada kapsamlı bir kaynak sağlamaktadır. Bu nedenle 2015 yılında gerçekleştirilen TIMSS sınavında 8. sınıf öğrencilerinin matematik başarısı ile ilişki gösteren değişkenlerin belirlenmesi politika yapıcılara, sivil toplum kuruluşlarına ve araştırmacılara ışık tutması bakımından önemli görülmektedir. Bu kapsamda bu araştırmada aşağıda yer alan araştırma sorularına yanıt aranmıştır:

Türkiye'de;

1.1. *öğrenci boyutu* altında yer alan;
- evdeki kaynaklar
- okul algısı
- duyuşsal alan

1.2. *öğretmen ve okul* boyutunda yer alan
- okulda başarıya verilen önem
- güvenli ve düzenli okul ortamı
- okul koşulları
- iş tatmini
- öğretmenin karşılaştığı sorunlar
- öğrencilerden kaynaklı sorunlar ile ilgili TIMSS 2015 puanları diğer ülke puanlarına kıyasla nasıldır?

2. Bu değişkenler öğrencilerin TIMSS 2015 matematik başarı puanlarını ne ölçüde yordamaktadır?

YÖNTEM

Çalışma Grubu

Bu çalışmanın verileri TIMSS 2015 uygulamasından elde edilmiştir. TIMSS 2015'de iki aşamalı tabakalı örnekleme yöntemi (stratified two-stage cluster sample design) kullanılmıştır. Birinci aşamada, okullar öğrenci sayılarına bağlı olasılıklarla seçilmiş, ikinci aşamada ise ilk aşamada seçilen okulların birer veya ikişer sınıfı rastgele seçim yöntemi ile belirlenmiştir (LaRoche, Joncas ve Foy, 2016). Mevcut çalışmada, bu yöntemler sonunda Türkiye'den TIMSS 2015'e seçilen tüm 8. sınıf öğrencileri ve sınıflarının (öğretmenlerin) verileri kullanılmıştır. Türkiye'den TIMSS 2015'e 6079 öğrenci (%48 kız, %52 erkek) ve 220 öğretmen (%47 kadın, %53 erkek) katılmıştır.

Veri Toplama Araçları

Bu çalışmanın ölçme araçlarını TIMSS 2015'te uygulanan matematik başarı testi, öğrenci ve öğretmen anketi oluşturmaktadır. Araştırmanın bağımlı değişkeni TIMSS matematik başarı testi ile elde edilen puanlardır. Araştırmanın bağımsız değişkenlerini TIMSS öğrenci ve öğretmen anketi sorularına verilen yanıtlardan elde edilen öğrenci, öğretmen ve okul özellikleri ile ilgili puanlar oluşturmaktadır. TIMSS 2015 öğrenci ve öğretmen anketinde matematik başarısı ile ilgili olduğu düşünülen öğrenci, öğretmen ve okul özelliklerini ölçmek amacı ile Likert tipi sorular kullanılmıştır. Bu anket soruları örtük yapıları ölçmek amacı ile sorulmuş, TIMSS uzmanları tarafından ConQuest 2.0 programı ve Madde Tepki Kuramı kullanılarak ölçülmek istenen yapıya ait ölçek puanları raporlanmıştır. Ölçek puanları ortalama 10, standart sapma 2 olacak şekilde hesaplanmıştır (Martin, vd., 2016).

Çalışmada öğrenci özellikleri öğrencilerin anket sorularına verdikleri yanıtlardan oluşturulmuştur. Ele alınan değişkenlerde öğrenci boyutu; *duyuşsal alan (öz-yeterlik, tutum ve öğrenme değeri), evdeki eğitimsel kaynaklar, okula aidiyet, zorbalık, öğretim etkinlikleri*, okul ve öğretmen boyutu; *okulda başarıya verilen önem, güvenli ve düzenli okul ortamı, okul koşulları, iş tatmini, öğretmenin karşılaştığı sorunlar, öğrencilerden kaynaklı sorunlar* ölçek puanları ile ölçülmüştür.

Duyuşsal Alanda yer alan öz-yeterlik, tutum ve öğrenme değerinin puanlanmasında 4'lü Likert tipi ölçek kullanılmıştır (tamamen katılmıyorum, katılmıyorum, katılıyorum, tamamen katılıyorum). *Öz-yeterlik* puanı (Students Confident in Mathematics) matematikte genelde iyiyimdir, matematiği hızla öğrenirim, matematikteki zor soruları yanıtlamada iyiyimdir gibi 9 madde kullanılarak oluşturulmuştur. Ölçekten alınan puan arttıkça öğrencinin öz-yeterlik inancı da artmaktadır. *Tutum* puanı (Students Like Learning Mathematics) matematik öğrenmekten hoşlanırım, matematikte pek çok ilginç şey öğrenirim, matematiği severim gibi 9 maddeye öğrencilerin katılma düzeyi kullanılarak hesaplanmıştır.Bu ölçekten alınan puanların artması öğrencilerin matematiğe yönelik tutumlarının yüksek olduğunu göstermektedir. *Öğrenme Değeri* puanı (Students Value Mathematics) matematik öğrenmek günlük yaşantımda bana yardım eder, üniversitede istediğim bölüme girmek için matematik öğrenmeliyim, istediğim işe girmek için matematikte iyi olmalıyım gibi 9 maddeye öğrencilerin katılma düzeyi kullanılarak hesaplanmıştır. Anketten alınan puanların yüksek olması öğrencilerin matematiğe yönelik öğrenme değerinin fazla olduğunu göstermektedir.

Evdeki Eğitimsel Kaynaklar puanı (Home Educational Resources) evdeki kitap sayısı, evde internet ve odaya sahip olma ile anne ve babanın eğitim düzeyi maddeleri kullanılarak oluşturulmuştur. Bu bilgiler kaç kitap olduğu, internete ve odaya sahip olup olmama ve anne babanın tamamladıkları eğitim düzeyi sorularak elde edilmiştir. Bu ölçek puanında yüksek puana sahip olma öğrencilerin daha fazla kaynağa sahip olduğunu göstermektedir.

Okula Aidiyet puanı (Students' Sense of School Belonging) okulda olmayı severim, okulda kendimi güvende hissederim, bu okula gitmekten gurur duyuyorum gibi 7 maddeye öğrencilerin katılım düzeyleri kullanılarak elde edilmiştir. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (tamamen katılmıyorum, katılmıyorum, katılıyorum, tamamen katılıyorum). Bu ölçekten öğrencilerin yüksek puana sahip olmaları kendilerini okula daha fazla ait hissettikleri anlamına gelmektedir.

Zorbalık puanı (Student Bullying) benimle dalga geçerler, beni oyunlarının dışında tutarlar, bana vururlar ve benzeri 8 maddeye öğrencilerin katılma düzeyi kullanılarak hesaplanmıştır. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (hiçbir zaman, yılda birkaç kez, ayda bir veya iki kere, haftada en az bir). Zorbalık puanının yüksek olması öğrencilerin daha az zorbalığa maruz kaldıklarını göstermektedir.

Öğretim Etkinlikleri puanı (Students' Views on Engaging Teaching in Mathematics Lessons) öğretmenimi anlamak kolaydır, öğretmenimin ne dediği ile ilgilenirim, öğretmenim yapmamız için ilginç şeyler verir, öğretmenimiz öğrenmemiz için çok çeşitli şeyler yapar gibi 10 madde kullanılarak oluşturulmuştur. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (tamamen katılmıyorum, katılmıyorum, katılıyorum, tamamen katılıyorum). Yüksek puan daha fazla ilgi ve sınıfta çeşitli şeyler yapıldığını göstermektedir.

Okul ve öğretmen boyutunda; *Okulda Başarıya Verilen Önem* puanı (School Emphasis on Academic Success) öğretmenlerin gözünden bir okulda öğretmenlerin öğretim programlarını uygulama düzeyi, öğretmenlerin öğrencilere örnek olma düzeyi, ailelerin öğrenci başarısı beklenti düzeyi ve öğrencilerin başarılı olma arzu düzeyi gibi 14 maddeye katılma düzeyi kullanılarak hesaplanmıştır. Puanlamada 5'li Likert tipi ölçek kullanılmıştır (çok düşük, düşük, orta, yüksek, çok yüksek). Bu ölçekte yüksek puan öğretmen algısına göre o okulda başarıya daha fazla önem verildiğini göstermektedir.

Güvenli ve Düzenli Okul Ortamı puanı (Safe and Orderly School) bu okul güvenli bir bölgededir, okulda kendimi güvende hissediyorum ve öğrenciler okul malzemelerini korur gibi 8 madde kullanılarak oluşturulmuştur. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (tamamen katılmıyorum, katılmıyorum, katılıyorum, tamamen katılıyorum). Bu ölçekte yüksek puan öğretmen algısına göre o okuldaki ortamın daha fazla güvenli olduğu anlamına gelmektedir.

Okul koşulları puanı (Problems with School Conditions and Resources) okulun ciddi bir bakıma ihtiyacı var, öğretmenlerin uygun bir çalışma alanı yok, öğretmenlerin yeteri kadar eğitim materyali yok gibi 7 madde öğretmenlerin katılma düzeyleri kullanılarak hesaplanmıştır. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (ciddi problem, orta düzey problem, küçük problem, problem değil).Yüksek puanı olan okullarda daha az problem yaşandığı belirtilmektedir.

İş Tatmini puanı (Teacher Job Satisfaction) öğretmenlik mesleğinden memnunum, bu okulda öğretmen olmaktan memnunum ve yaptığım iş ile gurur duyuyorum ve benzeri 7 madde kullanılarak oluşturulmuştur. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (hiçbir zaman, bazen, sıklıkla, çok sık). Bu ölçekte yüksek puanı olan bir öğretmenin iş tatmini seviyesi de yüksektir.

Öğretmenin Karşılaştığı Sorunlar puanı (Challenges Facing Teachers) sınıflarda çok fazla öğrenci var, çok fazla saat derse giriyorum, çok fazla idari işim var gibi 8 maddeye öğretmenlerin verdikleri yanıtlar kullanılarak hesaplanmıştır. Puanlamada 4'lü Likert tipi ölçek kullanılmıştır (tamamen katılmıyorum, katılmıyorum, katılıyorum, tamamen katılıyorum). Yüksek puana sahip öğretmenler daha az sorun algısına sahiptirler.

Öğrencilerden Kaynaklı Sorunlar puanı hesaplanırken (Teaching Limited by Student Needs) öğrencilerin ön bilgi ve becerilerde eksikleri var, dikkat dağıtıcı öğrenciler var ve ilgisiz öğrenciler var gibi 6 maddeye öğretmenlerin katılma düzeyleri kullanılmıştır. Puanlamada 3'lü Likert tipi ölçek kullanılmıştır (hiç, az, çok). Bu ölçekten alınan yüksek puan öğretmenlerin öğrencileri ile daha az sorun yaşadıkları anlamına gelmektedir.

Verilerin Analizi

Çalışmada ilk olarak öğrenci ve öğretmen anketindeki sorulara verilen yanıtlar kullanılarak elde edilen ölçek puanları ile ilgili güvenirlik katsayıları ve betimleyici istatistikler raporlanmıştır. Ölçek puanları ortalama 10 ve standart sapma 2 olacak şekilde hesaplandığı için ölçek puanlarına bakılarak Türkiye genelindeki durumun yanı sıra diğer ülkelere göre Türkiye'deki öğrenci ve öğretmen özellikleri de değerlendirilebilmektedir. Ortalaması 10'dan fazla olan ölçeklerde o özelliğin Türkiye'de daha fazla olduğu anlamına gelmektedir. Çalışmanın ikinci bölümünde ise öğrenci ve öğretmen özellikleri kullanılarak matematik başarısı tahmin edilmiştir. Bu amaçla, TIMSS matematik başarı puanları bağımlı değişken ve öğrenci ve öğretmen özellikleri değişkenleri bağımsız değişken olacak şekilde hiyerarşik regresyon analizi gerçekleştirilmiştir. TIMSS uygulamalarında kullanılan örneklem seçim yönteminde veriler öğrenciler ve bu öğrencilerin ilişkili olduğu öğretmenlerden toplandığı için hiyerarşik bir yapı bulunmakta ve yapılan analizlerde bu hiyerarşik yapıyı dikkate almak gerekmektedir. Bu çalışmadaki hiyerarşik regresyon analizleri TIMSS datasının özelliklerini dikkate alabilen MPLUS 7.4 programı ile yapılmıştır (Muthen ve Muthen, 2015).

Hiyerarşik regresyon analizinin temel varsayımları olan değişkenlerin normal dağılımı, bağımlı ve bağımsız değişkenler arasındaki doğrusallık ve hata varyanslarının bağımsızlığı incelenmiştir. Öğrenci düzeyinde bulunan 7 değişkenin ve sınıf düzeyinde bulunan 6 değişkenin normal dağılıp dağılmadığı çarpıklık ve basıklık değerleri ile incelenmiştir. Değerlerin genel olarak -1 ile +1 arasında değiştiği için (özyeterlik basıklık değeri 1.29; öğretmenin karşılaştığı sorunlar basıklık değeri 1.36;

öğrencilerden kaynaklı sorunlar basıklık değeri 1.18) normal dağılım varsayımından uzaklaşılmadığı görülmüştür. Bağımlı ve bağımsız değişkenler arasındaki doğrusallık varsayımı saçılma diyagramı ile incelendiğinde ise doğrusallığa tehdit oluşturan bir durum ile karşılaşılmamıştır. Hata varyanslarının bağımsızlığı varsayımı da ilgili saçılma diyagramı ile incelenmiş, hata varyansları arasında bir ilişki olmadığı görülmüştür.

BULGULAR

Ölçek Puanlarının Güvenirliği

TIMSS tarafından raporlanan ölçek puanlarının Türkiye için güvenirlik değerleri (Cronbach's Alpha) Tablo 1'de verilmiştir. Güvenirlik değerleri TIMSS tarafından raporlanmıştır (Martin, vd., 2016). Ortanca değerleri ise hesaplanarak tabloya eklenmiştir. Türkiye'deki öğrencilerin ve öğretmenlerin anket sorularına verdikleri yanıtların genel olarak tutarlı olduğu görülmektedir. George ve Mallery (2003) Cronbach's alpha değeri 0.90 civarı ise mükemmel, 0.80 civarı ise iyi, 0.70 civarı ise kabul edilebilir ve 0.60 civarı ise sorgulanmalıdır diye belirtmektedir. Öğrenci değişkenlerinde eğitimsel kaynaklar dışında tüm güvenirlik değerleri iyi veya mükemmel seviyededir. Eğitimsel kaynaklar değişkeninin güvenirlik değeri tüm ülkelerin en yüksek değeri olan 0.63 değerine oldukça yakın ve ortanca değerinden oldukça yüksektir. Bu ölçek kapsamında sadece 3 soru sorulmasının genel olarak düşük bir güvenirlik değeri çıkmasına sebep olduğu düşünülmektedir. Öğretmen değişkenlerinde tüm güvenirlik değerleri kabul edilebilir veya üzeri seviyededir. Genel olarak ölçek puanlarının tutarlı olduğu, Türkiye'den gelen yanıtların ise ortanca değerlerine yakın olduğu görülmektedir.

Tablo 1. Öğrenci, Öğretmen ve Okul Ölçek Puanlarının Güvenirlik Değerleri

	Türkiye	Ortanca Değeri	Uluslararası Minimum Değer	Uluslararası Maksimum Değer
Düzey-1 (Öğrenci)				
Öğrencilerin Evdeki Kaynakları				
Eğitimsel Kaynaklar	.62	.44	.27	.63
Öğrencilerin Okul Algısı				
Okula Aidiyet	.78	.82	.68	.88
Zorbalık	.81	.83	.74	.88
Öğretim Etkinlikleri	.89	.91	.86	.94
Öğrencilerin Duyuşsal Alanı				
Tutum	.92	.93	.87	.95
Öz-yeterlik	.87	.88	.68	.93
Öğrenme Değeri	.87	.88	.81	.91
Düzey-2 (Öğretmen ve Okul)				
Okulda Başarıya Verilen Önem	.89	.90	.82	.92
Güvenli ve Düzenli Okul Ortamı	.88	.86	.77	.90
Okul Koşulları	.88	.85	.78	.91
İş Tatmini	.88	.90	.85	.95
Öğretmenin Karşılaştığı Sorunlar	.72	.73	.50	.82
Öğrencilerden Kaynaklı Sorunlar	.71	.75	.61	.82

Birinci Alt Probleme İlişkin Bulgular

Betimleyici istatistikler

Bu bölümde çalışmada kullanılan öğrenci, öğretmen ve okul değişkenlerinin ortalama ve standart sapma değerleri verilmiştir. TIMSS tarafından raporlanan ölçek puanlarının kullanılmasının en önemli avantajı standart olan bu puanların genel ortalamalar ile karşılaştırmaya olanak sağlamasıdır. Ölçek puanları ortalama 10, standart sapma 2 olacak şekilde hesaplanmıştır (Martin, vd., 2016). Ortalamanın 10'dan fazla olması o özellikten Türkiye'de ortalamaya göre daha çok olduğu, 10'dan az olması da o özelliğin daha az olduğunu göstermektedir. Bu değerlerin istatistiksel olarak 10'dan farklı olup olmadığını değerlendirme amacı ile bu değerler tek örneklem t-testi ile kontrol edilmiştir. Fark bulunan değişkenler Tablo 2'de gösterilmektedir.

Tablo 2. Öğrenci, Öğretmen ve Okul Ölçek Puanlarının Betimleyici İstatistik Değerleri

	Ort.	SS
Düzey-1 (Öğrenci)		
Öğrencilerin Evdeki Kaynakları		
Eğitimsel Kaynaklar	9.12***	1.93
Öğrencilerin Okul Algısı		
Okula Aidiyet	10.60***	2.00
Zorbalık	10.28***	1.97
Öğretim Etkinlikleri	10.57***	1.84
Öğrencilerin *Duyuşsal Alanı*		
Tutum	10.26***	1.97
Öz-yeterlik	9.75***	2.28
Öğrenme Değeri	10.05	2.09
Düzey-2 (Öğretmen ve Okul)		
Okulda Başarıya Verilen Önem	9.19***	1.92
Güvenli ve Düzenli Okul Ortamı	9.28***	2.15
Okul Koşulları	8.94***	2.19
İş Tatmini	9.77**	1.81
Öğretmenin Karşılaştığı Sorunlar	11.64***	2.20
Öğrencilerden Kaynaklı Sorunlar	8.74***	1.63

*$p < .05$. **$p < .01$. ***$p < .001$.

Tablo 2 incelendiğinde, Türkiye'deki öğrencilerin evde sahip oldukları eğitimsel kaynakların diğer ülkelerdeki öğrencilerin sahip oldukları kaynaklardan daha az olduğu görülmektedir. Türkiye'deki öğrenciler kendilerini okullarına daha fazla ait hissetmekte ve diğer ülkelerin öğrencilerinden daha az zorbalığa maruz kalmaktadırlar. Öğretim etkinlikleri hakkında Türkiye'deki öğrencilerin daha olumlu değerlendirmeler yaptıkları görülmektedir. Türkiye'deki öğrenciler diğer ülkelerdeki öğrencilere göre matematik dersi ile ilgili olarak daha olumlu tutuma sahipken, öz-yeterlik bakımından ise kendilerini daha az yeterli hissetmektedirler. Matematiğe verilen önem değerlendirildiğinde, Türkiye'deki ve diğer ülkelerdeki öğrencilerin ortalamaları arasında anlamlı bir fark olmadığı görülmektedir.

Tablo 2'de yer alan öğretmen ve okul özelliklerine göre okullarda başarıya verilen önem diğer ülkelere göre daha düşük seviyededir. Türkiye'deki öğretmenlerin okullarını güvenli ve düzenli görme oranları diğer ülkelerin ortalamalarının altındadır. Okul koşulları incelendiğinde ise Türkiye'deki öğretmenlerin okullarında diğer ülke öğretmenlerine göre daha az sorun olduğunu belirtmişlerdir. Türkiye'deki öğretmenlerin yaptıkları meslekten daha az tatmin oldukları görülmektedir. Öğretmenin karşılaştığı sorunlara bakıldığında Türkiye'deki öğretmenler sınıflarında çok fazla öğrenci olması veya çok fazla saat derse girme gibi durumları daha az yaşadıklarını belirtirken, öğrencilerin ön bilgi ve becerilerinde eksik olma, dikkat dağıtıcı ve ilgisiz öğrencilerin var olması gibi sorunlardan daha fazla şikâyetçi olduklarını belirtmişlerdir.

İkinci Alt Probleme İlişkin Bulgular

Matematik başarısını tahmin edebilen öğrenci, öğretmen ve okul özellikleri

Bu kısımda Türkiye'deki öğrencilerin matematik başarısını yordayan öğrenci, öğretmen ve okul özellikleri raporlanmaktadır (Tablo 3). Elde edilen sonuçlara göre matematik başarısını yordamada en önemli öğrenci özelliği öz-yeterliktir (β = .56). Öğrencilerin öz-yeterlik değerleri arttıkça, matematik başarıları da artmaktadır. Matematik başarısını tahmin etmede önemli rolü olan bir başka değişken öğrencilerin evde sahip oldukları eğitimsel kaynaklardır (β = .18). Öğrencilerin evlerinde sahip oldukları eğitimsel kaynaklar arttıkça, matematik başarıları da artmaktadır. Bu çalışmada başarıyı açıklayan faktörlerden en önemli iki tanesi olan öz-yeterlik ve evde sahip olunan eğitimsel kaynakların Türkiye'deki öğrenciler için uluslararası ortalamanın altında olduğunu belirtmek gerekir. Öğrencilerin matematik dersine karşı tutumları ve matematik başarısı arasında negatif bir ilişki bulunmuştur (β = -.10). Bu negatif ilişki değişkenler bir arada ele alındığında öğrencilerin öz-yeterlik gibi diğer değişkenleri sabit tutulduğunda tutum ve matematik başarısı arasında negatif bir ilişki olduğu anlamına gelmektedir. Matematik başarısı ile anlamlı bir ilişkisi olan ama göreli olarak tahminde daha az önemi olan diğer değişkenler zorbalık (β = .07), okula aidiyet (β = -.06) ve öğretim etkinlikleridir (β = .05). Bu bulgular daha az zorbalığa maruz kalan öğrencilerin daha başarılı, okula kendilerini daha fazla ait hisseden öğrencilerin daha az başarılı ve daha etkin öğretim yaşayan öğrencilerin daha başarılı olduğu anlamına gelmektedir. Bu çalışmada öğrencilerin matematiğe verdiği önem ve matematik başarısı arasında istatistiksel olarak anlamlı bir ilişki bulunamamıştır. Araştırmada ele alınan öğrenci özellikleri öğrenciler arasındaki matematik başarısı farklılıklarının %34'ünü açıklayabilmektedir.

Tablo 3. Matematik Başarısını Yordayan Öğrenci ve Öğretmen Özellikleri

	Standart Olmayan Katsayılar	Standart Katsayılar
Düzey-1 (Öğrenci)		
Öğrencilerin Evdeki Kaynakları		
Eğitimsel Kaynaklar	8.29***	.18***
Öğrencilerin Okul Algısı		
Okula Aidiyet	-2.73***	-.06***
Zorbalık	3.03***	.07***
Öğretim Etkinlikleri	2.21**	.05**
Öğrencilerin Duyuşsal Alanı		
Tutum	-3.99***	-.10***
Öz-yeterlik	21.00***	.56***
Öğrenme Değeri	-.49	-.01
Düzey-2 (Öğretmen)		
Okulda Başarıya Verilen Önem	9.85***	.40***
Güvenli ve Düzenli Okul Ortamı	-1.05	-.05
Okul Koşulları	1.15	.05
İş Tatmini	-.02	.00
Öğretmenin Karşılaştığı Sorunlar	-.99	-.04
Öğrencilerden Kaynaklı Sorunlar	6.88**	.24**
Okullar içi açıklanan varyans	%34	
Okullar arası açıklanan varyans	%29	

*$p < .05$. **$p < .01$. ***$p < .001$.

Tablo 3'te matematik başarısı ile ilişki gösteren öğretmen ve okul özellikleri incelendiğinde okulda başarıya verilen önem (β = .40) ve öğrencilerden kaynaklı sorunlar (β = .24) değişkenlerinin matematik başarısını tahmin edebildikleri görülmektedir. Matematik başarısını tahminde rolü olan bir diğer değişken öğrencilerden kaynaklı sorunlardır. Öğrencileri daha az sorunlara sahip olan veya bu şekilde düşünen öğretmenlerin sınıflarında bulunan öğrenciler daha başarılıdır. Buna karşın, Türkiye'deki öğretmenler diğer ülkelerdeki öğretmenlere göre daha fazla bu konuda sorun yaşadıklarını raporlamışlardır. Bu çalışmanın öğretmenler tarafından raporlanan diğer değişkenleri olan güvenli ve düzenli okul ortamı, okul koşulları, iş tatmini ve öğretmenin karşılaştığı sorunlar ile matematik başarısı arasında bir ilişki bulunamamıştır. Araştırmada ele alınan okul ve sınıf ile ilgili faktörler bir araya geldiklerinde sınıflar arasındaki matematik başarısı farklılıklarının %29'unu açıklayabilmektedir.

SONUÇLAR ve TARTIŞMA

TIMSS, PISA ve PIRLS gibi uluslararası sınavlar, her katılımcı ülkeye başarı sonuçlarını yorumlamak, eğitim ve öğretim programları uygulamalarında meydana gelen değişiklikleri izlemek için kapsamlı bir kaynak sağlamaktadır (Mullis, Martin, Ruddock, O'Sullivan ve Preuschoff, 2009). Özellikle Türkiye gibi eğitim sisteminde sorunları olan ve bu sorunlara yönelik çözüm arayışı içerisinde olan ülkelere uluslararası düzeyde yapılan sınavlara ait sonuçlar önemli veri kaynakları sunmaktadır (Ölçüoğlu, 2015). Yapılan bu sınavlarda; fen, matematik, okuma gibi temel alanlarla birlikte öğrencilere uygulanan anketlerle motivasyonları, kendileri hakkındaki düşünceleri, öğrenme süreçlerine ilişkin psikolojik özellikleri, öğrenim gördükleri okul ortamları ve aileleri ile ilgili veriler toplanmaktadır. Bu veriler, öğrencilerin bilişsel başarılarına yönelik elde edilen verilerin yorumlanmasında kullanılmaktadır (MEB, 2016). Bu araştırmada 2015 yılında gerçekleştirilen TIMSS sınavına katılan 8. sınıf öğrencilerinin matematik dersine yönelik bilişsel alana ait akademik başarıları ile öğrenci, öğretmen ve okul değişkenlerinin nasıl bir ilişki içinde olduğunun belirlenmesi amaçlanmıştır. Araştırmada öğrenci faktörü altında yer alan bağımsız değişkenler öğrencilerin duyuşsal alan (tutum, öz-yeterlik ve öğrenme değeri), evdeki kaynakları, okula aidiyet, zorbalık, öğretim etkinlikleridir. Çevre (okul ve öğretmen) faktörü altında yer alan değişkenler ise okulda başarıya verilen önem, güvenli ve düzenli okul ortamı, okul koşulları, iş tatmini, öğretmenin karşılaştığı sorunlar ve öğrencilerden kaynaklı sorunlardır.

Bu araştırma kapsamında yukarıda belirtilen öğrenci değişkenlerinin TIMSS 2015'teki matematik başarısındaki farklılıkların %34'ünü açıkladığı görülmektedir. Bu faktörlerden duyuşsal alan boyutunda yer alan öz-yeterlik 8. sınıf öğrencilerinin TIMSS 2015'te matematik başarılarını yordamada en önemli değişkendir. Diğer bir ifadeyle öz-yeterlik inançları yüksek olan öğrencilerin matematik başarıları daha yüksektir. Benzer şekilde TIMSS 1999, 2007 ve 2011 yıllarında Türk öğrencilerin öz-yeterlik inancı matematik başarılarını açıklamada önemli bir değişken olarak bulunmuştur (Demir ve Kılıç, 2010; Doğan ve Barış, 2010; Yavuz vd., 2017). Bu açıdan düşünüldüğünde ailede ve okulda öğrencilerin öz-yeterlik düzeylerini yüksek tutacak ortamların sağlanması öğrencilerin matematik başarılarının artmasında katkı sağlayacağı söylenebilir. Çünkü yüksek öz-yeterliğe sahip öğrencilerin kendilerine verilen görevi tamamlama eğilimlerinin yüksek olduğu, herhangi bir zorlukla karşı karşıya geldiklerinde sebat ederek daha çok çalıştıkları görülmektedir (Pajares, 2008). Öz-yeterlik bireylerin pozitif ve gerçekçi bakış açısıyla beklentilerini ve yeteneklerini yönetebilme fırsatı sağladığından bu inanç öğrenilebilir ve inşa edilebilir (Ker, 2016). Bu sebepten dolayı öğrencilerin matematik başarılarının artırılması için öğretmenlerin öğrencilerinin matematik öz-yeterliklerini geliştirmeleri gerekmektedir. Başka bir ifadeyle eksik öz-yeterlik algısı eksik yetenek anlamına geldiğini düşünürsek öğretmenlerin öğrencilerinin güvenlerini güçlendirmeleri ve matematiksel yeterliklerini inşa etmeleri önemlidir (Chen, 2014).

Duyuşsal alanla ilgili elde edilen bir diğer bulgu, tutum değişkeniyle ilgilidir. Öğrencilerin matematik dersine karşı tutumları ile matematik başarı puanları arasında negatif yönde ve anlamlı bir ilişki çıkmıştır. Öğrencilerin matematiğe yönelik tutum puanları arttıkça matematik başarıları düşmektedir. Tutum değişkeniyle ilgili elde edilen bu bulgu araştırmanın çarpıcı bulguları arasında yer almaktadır. Çünkü öğrenme ile ilgili en kritik öneme sahip yapının matematiğe karşı tutum olduğu (Ölçüoğlu ve Çetin, 2016) ve matematikteki başarısızlığının sebepleri arasında öğrencilerin matematiğe yönelik olumsuz tutumları gösterilmektedir (Baykul, 2009). Türk öğrencilerinin daha önce TIMSS sınavlarındaki tutum değişkenine yönelik sonuçları da farklılık göstermektedir. Örneğin, Türk öğrencilerinin matematiğe yönelik tutum puanları ile TIMSS 2007 matematik başarıları arasında negatif ilişki görülürken (Şişman, Acat, Aypay ve Karadağ, 2011), TIMSS 1999'da tutum değişkeninin matematik başarı puanları üzerinde anlamlı bir etkisi bulunamamıştır (Doğan ve Barış, 2010; Uzun, Bütüner ve Yiğit, 2010). Türk öğrencilerinin matematiğe karşı nispeten olumsuz tutum sergilemesinin nedeni arasında Türkiye'deki rekabetçi sınav sistemi gösterilebilir. Çünkü rekabetçi sınav sisteminin ve öğretmenlerin öğrencileri teşvik eksikliğinin öğrencilerin matematik dersine karşı olumsuz tepki geliştirilebilecekleri ifade edilmektedir (Leung, 2002).

Duyuşsal alan içerisinde yer alan diğer bir değişkene ait bulguya bakıldığında; matematiğe verilen önem ile öğrencilerin matematik başarıları arasında anlamlı bir ilişki çıkmamıştır. Benzer şekilde 2007 ve 2011 yıllarında TIMSS uygulamalarına Türkiye'den katılan 8. sınıf öğrencilerinin matematik başarı puanları ile matematiğe verilen değer puanları arasındaki ilişki her iki uygulama dönemi için de anlamlı olmadığı bulunmuştur (Arıkan vd., 2016; Yavuz vd., 2017). Matematiğe verilen değer ile matematik başarı arasında bir ilişkinin ortaya çıkmaması matematiğe değer veren öğrencilerle, değer vermeyen öğrenciler arasında bir başarı farkı olmadığı şeklinde yorumlanabilir. Aynı zamanda öğrencilerin yaşamın her alanın yer alan matematiğin öneminin farkına varması da bu duruma neden olduğu söylenebilir (Yavuz vd., 2017).

TIMSS 2015 sınavında matematik başarısını tahmin etmede öz-yeterlik inancından sonra önemli rolü olan bir başka değişken öğrencilerin evde sahip oldukları eğitimsel kaynaklardır. TIMSS kapsamında evdeki kitap sayısı, bilgisayar ve eğitim ile ilgili bilgisayar programı ve internet, odaya sahip olma, anne ve babanın eğitim düzeyi gibi durumlar evdeki eğitimsel kaynaklar olarak ele alınmaktadır. Bu araştırma sonucuna göre; öğrencilerin evlerinde sahip oldukları eğitimsel kaynaklar arttıkça, matematik başarıları da artmaktadır. Başka bir ifade ile anne-babanın eğitim düzeyi yükseldikçe, evde bulunan kitap sayısı ve eğitime destek verecek diğer araç-gereçler (bilgisayar, internet vb.) arttıkça matematik başarısı artıyor denilebilir. Çünkü daha fazla kaynaklara sahip olan ailelerden gelen öğrenciler ayrıcalıklı okullara gönderilmekte, daha iyi öğretmenlerden eğitim almakta ve yüksek akademik beklentileri olan ortamlarda yer almaktadır (Chiu, 2010). Yine anne-babanın eğitim düzeyi öğrenciye verilen desteğin (kitap, evde birlikte çalışma vb.) artmasına sebep olabilmektedir. Bu durumda sosyo-ekonomik düzey (Oral ve Mcgivney, 2013) ve kültür eğitimde önemli bir faktör olarak karşımıza çıkmaktadır. Yapılan araştırmalar incelendiğinde bu araştırma ile benzer bulguların ortaya çıktığı görülmektedir. Örneğin; TIMSS 1999 (Akyüz ve Berberoğlu, 2010; Yayan ve Berberoğlu, 2004), TIMSS 2011 (Akyüz, 2014; Bayar ve Bayar, 2013; Kılıç ve Aşkın, 2013; Ölçüoğlu ve Çetin, 2016), PISA 2006 (Özer ve Anıl, 2011) ve PISA 2012 (Usta, 2014) verileriyle yapılan araştırmalarda öğrencilerin evdeki eğitimsel kaynakların matematik akademik başarısıyla anlamlı düzeyde bir ilişki içerisinde olduğu ve matematik başarısını yordayan önemli değişkenlerden biri olduğu ortaya çıkmıştır.

Araştırmada matematik başarısı ile anlamlı bir ilişkisi olan ama göreli olarak tahminde daha az önemi olan diğer değişkenler zorbalık, okula aidiyet ve öğretim etkinlikleridir. Öğrencilerin okul algısı ile ilgili değişkenlere (zorbalık, okula aidiyet ve öğretim etkinlikleri) ait bulguları incelendiğinde; okullarda daha az zorbalığa maruz kalan öğrencilerin daha başarılı, okula kendilerini daha fazla ait hisseden öğrencilerin daha az başarılı ve daha etkin öğretim yaşayan öğrencilerin daha başarılı olduğu görülmüştür. Araştırmadan elde edilen bu bulgular alanyazın ile paralellik göstermektedir. TIMSS 2011 değişkenleri arasında yer alan şiddet ve zorbalık olaylarının diğer ülkelere oranla Türkiye'de daha fazla görüldüğü ve bu durumun öğrencilerin matematik başarısını olumsuz yönde etkilediği belirtilmektedir (Buluç, 2014; Yavuz vd., 2017). Benzer şekilde TIMSS 2007'de okullarda zorbalığa maruz kalan öğrencilerin akademik başarıları azaldığı görülmektedir (Yavuz vd., 2017). Elde edilen bu bulgular yüksek zorbalığa maruz kalma düzeyine sahip olan okullara devam eden öğrencilerin okuldaki performanslarının daha düşük olabileceğini göstermektedir (Strøm vd., 2013). Bu nedenle okul ortamında yaşanan zorbalık olaylarının yakından takip edilmesi ve bu olayların önlenmesi için gerekli tedbirlerin alınması öğrencilerin ruhsal sağlıkları ve kişilik gelişimleri kadar öğrenme ortamlarında sağlanan öğretimin kalitesi açısından da büyük öneme sahiptir (Yıldırım, Yıldırım, Yetişir ve Ceylan, 2013). Güvenli ve düzenli olmayan bir okul ortamında yaşayan ve kendini yaşadığı okula ait hissetmeyen öğrenciler okulun önemini ve değerini anlayamamakta; okulu can sıkıcı, rutin ve ileriki yaşamlarında hiçbir işe yaramayacak bir yer olarak görmelerine neden olabilmektedir (Şimşek ve Katıtaş, 2014). Öğrencilerin olumsuz, hoş olmayan bir iklime sahip veya kendilerini dışlanmış hissettikleri bir okula gitmek istememeleri (Özdemir, Sezgin, Şirin, Karip ve Erkan, 2010) onların akademik başarılarına da olumsuz etki yapabileceği söylenebilir. Yapılan araştırmalar bu bulguyu desteklemektedir (Goodenow ve Grady, 1993; Sarı, 2012; 2013). Yine öğrencilerin kendilerini okula ya da sınıfa ait hissetmemeleri öğretmenlerin sınıf öğretim etkinliklerinden kaynaklanabilir. Öğretmenin öğretim sürecine öğrenciyi sürece dâhil edememesi ve onun ilgi ve yeteneklerini göz önünde bulunduramaması öğrencilerin öğretmenle olan sınıf içi iletişimine ve dolayısıyla akademik başarısına etki edebilir. Matematik öğretim programındaki yeni yaklaşımlar ve bu bağlamda uygulanan programlar öğrencilerin matematiği anlamlı öğrenmesine, problem çözme becerilerini kazandırmasına ve becerilerin günlük hayatla olan ilişkisi üzerine yoğunlaşmaktadır. Bu sebepten dolayı öğretmenlerin sınıf içi öğretim uygulamaları ve öğrencilerin bu öğretim uygulamalarına yönelik algı ve anlayışları önem kazanmaktadır. İlgi, istek ve yeteneklerinin farkında olunan, olumlu bir iletişim süreci içinde olunan bir öğretim süreci bireyin sadece bilişsel özelliklerini değil duyuşsal özelliklerini de olumlu bir şekilde etkileyecektir.

Öğrenmede çevre faktörü altında ele alınan okulda başarıya verilen önem, güvenli ve düzenli okul ortamı, okul koşulları, iş tatmini, öğretmenin karşılaştığı sorunlar ve öğrencilerden kaynaklı sorunların hepsi bir araya geldiğinde öğrencilerin matematik başarılarının %29'unu açıklamaktadır. Öğrenmede çevre faktöründe yer alan okulda başarıya verilen önem ve öğrencilerden kaynaklı sorunlar öğrencilerin matematik başarısını yordamada iki önemli değişken olmuştur. Bir başka ifadeyle daha fazla başarıya önem verilen okullarda öğrenim gören öğrencilerin daha başarılı oldukları ve öğrencileri daha az sorunlara sahip olan veya bu şekilde düşünen öğretmenlerin sınıflarında bulunan öğrenciler daha başarılı oldukları söylenebilir. Her ne kadar çevresel faktörlerden okulda başarıya verilen önem ve öğrencilerden kaynaklı sorunlar öğrencilerin matematik başarısını yordamada önemli görülse de Türkiye'nin okulda başarıya verilen önem ortalaması diğer ülkelerin gerisinde olduğu ve öğretmenler diğer ülkelerdeki öğretmenlere göre daha fazla bu konuda sorun yaşadıkları görülmektedir (Martin vd., 2016).

Araştırmanın çevresel faktörleri içerisinde yer alan güvenli ve düzenli okul ortamı, okul koşulları, iş tatmini ve öğretmenin karşılaştığı sorunlar ile matematik başarısı arasında bir ilişki bulunamamıştır. Bu açıdan bakıldığında araştırmada güvenli ve düzenli okul ortamı ve okul koşulları ile öğrencilerin matematik başarıları arasında anlamlı bir ilişkinin ortaya çıkmaması araştırmanın bir diğer çarpıcı bulguları arasında yer almaktadır. Çünkü güvenli ve düzenli ortama sahip okulların öğrencilerin akademik başarısını olumlu yönde etkilediği belirtilmektedir (Abazaoğlu, Yatağan, Yıldızhan, Arifoğlu ve Umurhan, 2015; Buluç, 2014). Öğrencinin okulun sağlamış olduğu güvenilir ortamdan dolayı yaşadığı rahatlık onun okula yönelik motivasyonunu (Yaman vd., 2010) ve dolayısıyla bu da başarısını etkileyebileceği söylenebilir. Araştırmadan elde edilen bu bulgu alanyazın araştırmalarıyla paralellik göstermemektedir. Örneğin, TIMSS ve PIRLS 2011 sınavlarına ait sonuçlara bakıldığında düzensiz bir çevreye sahip ve okulda şiddetin fazla olduğu ortamlarda eğitim gören öğrencilerin düzenli ve güvenli öğrenme ortamlarda eğitim gören öğrencilere oranla daha düşük başarılı olduğu görülmektedir (Aydın, 2015).

Sonuç olarak; TIMSS sonuçları göz önünde bulundurulduğunda; matematik akademik başarısı ile yüksek ilişkisi bulunan birey ve çevre değişkenlerinin öğrencinin öğrenmesinde önemli farklılık sağlamaktadır. Bu sebepten dolayı öğretmenlerin öğrencilerin matematik dersine yönelik inanç, yeterlik algılarını ve tutumlarını da artırıcı motivasyonel stratejileri kullanmaları, sınıf içi öğretim uygulamalarının matematiği anlamlandırmada daha etkili yöntem, teknik ve stratejilerinin kullanılması gerekmektedir. Ayrıca sınıf ve okul ortamlarının öğretim sürecini destekleyici olması noktasında gerekli ortamın ve kaynakların sağlanması gerekmektedir. Öğrencilerin kendilerini okula ait ve okulda güvende hissetmeleri için gerekli adımların okul yöneticileri, aileler ve öğretmenlerle birlikte işbirliği içinde uygulanması gerekmektedir. Sosyo-ekonomik düzeyi düşük olan ailelerin bulunduğu çevrelerde okulların daha çok kaynaklara sahip olmasına dikkat edilmeli ve bu okullardaki çocukların okul kaynaklarından daha fazla yararlanmaları için okulda kalma sürelerinin uzatılması sağlanabilmelidir. Bu düşünceyi merkeze alan ülkelerde farklı okul modelleri geliştirilmiş olup bu gibi okulların uygulamaları incelenerek yeni okul modelleri geliştirilebilir. Bu sayede yeterli kaynakların olmamasından doğan imkânsızlıklar minimize edilmeye çalışabilir.

KAYNAKÇA

Abazaoğlu, İ., Yatağan, M., Yıldızhan, Y., Arifoğlu, A. ve Umurhan, H. (2015). Öğrencilerin matematik başarısının uluslararası fen ve matematik eğilimleri araştırması sonuçlarına göre değerlendirilmesi. *Turkish Studies-International Periodical for the Languages, Literature and History of Turkish or Turkic Volume*, *10*(7), 33-50. doi: 10.7827/TurkishStudies.7781

Abazaoğlu, İ. ve Aztekin, S. (2015). *Öğretmen moral ve motivasyonlarının öğrencilerin fen ve matematik başarılarına etkisi (Singapur, Japonya, Finlandiya ve Türkiye).* Uluslararası Eğitim Kongresi: Gelecek için Eğitim, Ankara, Türkiye.

Akyüz, G. (2014). TIMSS 2011'de öğrenci ve okul faktörlerinin matematik başarısına etkisi. *Eğitim ve Bilim, 39*(172), 150-162.

Akyuz, G., & Berberoglu, G. (2010). Teacher and classroom characteristics and their relations to mathematics achievement of the students in the TIMSS. *New Horizons in Education, 58*(1), 77-95.

American Federation of Teachers. (2006). *Building minds, minding buildings: Turning crumbling schools into environments for learning*. Retrieved from: http://www.chicagoacts.org/storage/documents/minding-bldgs.pdf.

Anderson, C. M. (2010). *Linking perceptions of school belonging to academic motivation and academic achievement amongst student athletes: A comparative study between high-revenue student athletes and non-revenue student athletes* (Doctoral dissertation, University of California). Retrieved from: https://escholarship.org/uc/item/8nt3g57h.

Arıkan, S., van de Vijver, F. J. R., & Yağmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian students in TIMSS 2007 and 2011. *EURASIA Journal of Mathematics, Science and Technology Education*, *12*(8), 2039-2059. doi: 10.12973/eurasia.2016.1268a

Ayan, A. (2014). *Ortaokul öğrencilerinin matematik öz-yeterlik algıları, motivasyonları, kaygıları ve tutumları arasındaki ilişki* (Yüksek lisans tezi, Balıkesir Üniversitesi, Fen Bilimleri Enstitüsü, Balıkesir). https://tez.yok.gov.tr/UlusalTezMerkezi adresinden edinilmiştir.

Aydın, M. (2015). *Öğrenci ve okul kaynaklı faktörlerin TIMSS matematik başarısına etkisi.* (Doktora tezi, Necmettin Erbakan Üniversitesi, Eğitim Bilimleri Enstitüsü, Konya). https://tez.yok.gov.tr/UlusalTezMerkezi adresinden edinilmiştir.

Aydın, Y. (1993). Matematik öğretmeni nasıl yetiştirilmeli. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 9*, 109–114.

Baker, L., & Bernstein, H. (2012). *The impact of school buildings on student health and performance: A call for research.* Retrieved from http://www.centerforgreenschools.org/sites/default/files/resource-files/McGrawHill_ImpactOnHealth.pdf .

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bayar, V. ve Bayar, S. A. (2013). *TIMSS 2011 matematik başarısı ulusal değerlendirme raporu.* Türk Eğitim Sendikası TIMSS 2011 Matematik Başarısı Ulusal Değerlendirme Raporu, Ankara. https://www.turkegitimsen.org.tr/upload_doc/00_2012_y/00_yok/TIMSS.docx adresinden edinilmiştir.

Baykul, Y. (2009). *İlköğretimde matematik öğretimi: 1-5. sınıflar için*. Ankara: Pegem Akademi.

Bayrakçı, M. (2007). Sosyal öğrenme kuramı ve eğitimde uygulanması. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi*, 14, 198-210.

Bilican, S., Demirtaşlı, R. N. ve Kilmen, S. (2011). Matematik dersine ilişkin türk öğrencilerin tutum ve görüşleri: TIMSS 1999 ve TIMSS 2007 karşılaştırması. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*(3), 1277-1283.

Bloom, B. S. (2012). *İnsan nitelikleri ve okulda öğrenme* (Çev. D. A. Özçelik). Ankara: Pegem Akademi.

Brese, F., & Mirazchiyski, P. (2010, July). *Measuring students' family background in large-scale education studies.* 4th IEA International Research Conference, Gothenburg, Sweden.

Buluç, B. (2014). TIMSS 2011 sonuçları çerçevesinde, okul iklimi değişkenine göre öğrencilerin matematik başarı puanlarının analizi. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, *33*, 105-121.

Büyükkaragöz, S. ve Çivi, C. (1999). *Genel öğretim metodları* (10. Baskı). İstanbul: Beta.

Clark, H. (2002). *Building education: The role of the physical environment in enhancing teaching and research. Issues in practice.* ERIC Document Number: 472 377.

Çalık, T., Kurt, T. ve Çalık, C. (2011). Güvenli okulun oluşturulmasında okul iklimi: Kavramsal bir çözümleme. *Pegem Eğitim ve Öğretim Dergisi*, *1*(4), 73-84.

Çelik E. (2012). *Matematik problemi çözme başarısı ile üstbilişsel özdüzenleme, matematik özyeterlik ve özdeğerlendirme kararlarının doğruluğu arasındaki ilişkinin incelenmesi* (Doktora tezi, Marmara Üniversitesi, Eğitim Bilimleri Enstitüsü, İstanbul). https://tez.yok.gov.tr/UlusalTezMerkezi adresinden edinilmiştir

Chen, Q. (2014). Using TIMSS 2007 data to build mathematics achievement model of fourth graders in Hong Kong and Singapore. *International Journal of Science and Mathematics Education, 12*, 1519–1545.

Chen, P. P. (2003). Exploring the accuracy and the predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*, 79-92.

Chiu, M. M. (2010). Effects of inequality, family and school on mathematics achievement: Country and student differences. *Social Forces, 88*(4), 1645-1676.

Danielson, C. (2002). E*nhancing student achievement: A framework for school improvement*. Association for Supervision ve Curriculum Development, USA, Alexandria, VA.

Doğan, N. ve Barış, F. (2010). Tutum, değer ve özyeterlik değişkenlerinin TIMSS-1999 ve TIMSS-2007 sınavlarında öğrencilerin matematik başarılarını yordama düzeyleri. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1), 44-50.

Demir, İ. ve Kılıç, S. (2010). Öğrencilerin matematiğe karşı tutumlarının öğrenci başarısına etkisi. *İstanbul Aydın Üniversitesi Dergisi, 2*(5), 50-70.

Demir, M. K. ve Arı, E. (2013). Öğretmen sorunları-Çanakkale ili örneği. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi*, *32*(1), 107-126.

Duru, E. ve Balkıs, M. (2015). Birey-çevre uyumu, aidiyet duygusu, akademik doyum ve akademik başarı arasındaki ilişkilerin analizi. *Ege Eğitim Dergisi*, *16*(1), 122-141.

Ekinci, A. (2014). İlköğretim okullarında çalışan müdür ve öğretmenlerin mesleki sorunlarına ilişkin görüşleri. *İlköğretim Online*, *9*(2), 734-748.

Ekizoğlu, N. ve Tezer, M. (2007). İlköğretim öğrencilerinin matematik dersine yönelik tutumları ile matematik başarı puanları arasındaki ilişki. *Cypriot Journal of Educational Sciences*, *2*(1), 43-57.

Engin, A. O., Özen, Ş. ve Bayoğlu, V. (2009). Öğrencilerin okul öğrenme başarılarını etkileyen bazı temel değişkenler. *Sosyal Bilimler Enstitüsü Dergisi*, *3*, 125-156.

Erdoğan, İ (1996). *İşletme yönetiminde örgütsel davranış*. İstanbul: Avcıoğlu.

Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. *Learning and Individual Differences, 19*(4), 499-505.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction, 17*(5), 478-493.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* 11.0 update (4th ed.). Boston: Allyn ve Bacon.

Goodenow, C., & Grady, K. E. (1993). The relationship of school belonging and friends’ values to academic motivation among urban adolescent students. The *Journal of Experimental Education*, *62*(1), 60-71.

İksara. (2013). *Okul güvenliği araştırması.* http://content.bahcesehir.edu.tr/public/files/files/CSG_Okul_V5.pdf adresinden edinilmiştir.

Jan, A., & Husain, S. (2015). Bullying in elementary schools: Its causes and effects on students. *Journal of Education and Practice, 6*(19), 43-56.

Karacaoğlu, Ö. C., & Kaçar, E. (2010). Yenilenen programların uygulanmasında öğretmenlerin karşılaştığı sorunlar. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 7*(1), 45-58.

Karasolak, K. ve Sarı, M. (2011). Mimarî özellikleri farklı okullardaki öğrenci ve öğretmenlerin okullarının binası hakkındaki görüşlerinin incelenmesi. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 3*(40), 132-154.

Keçeli-Kaysılı, B. (2008). Akademik başarının arttırılmasında aile katılımı. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi*, *9*(1), 69-83.

Ker, H. W. (2016). The effects of motivational constructs and engagements on mathematics achievements: a comparative study using TIMSS 2011 data of Chinese Taipei, Singapore, and the USA. *Asia Pacific Journal of Education*, *37*(2), 135-149. doi: 10.1080/02188791.2016.1216826

Kılıç, S., & Aşkın, Ö. E. (2013). Parental influence on students’ mathematics achievement: The comparative study of Turkey and best performer countries in TIMSS 2011. *Procedia - Social and Behavioral Sciences*, *106*, 2000-2007. doi: 10.1016/j.sbspro.2013.12.228

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 3.1-3.37). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-3.html adresinden edinilmiştir.

Lee Van Horn, M. (2003). Assessing the unit of measurement for school climate through psychometric and outcome Analyses of the school climate survey. *Educational and Psychological Measurement, 63*(6), 1002-1019.

Leder, G. C., & Forgasz, H. J. (2006). Affect and mathematics education: PME perspectives. In A. Gutiérrez, & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education: Past, present and future* (1st ed., 403-427). Rotterdam, The Netherlands: Sense Publishers.

Leung, F. K. (2002). Behind the high achievement of East Asian students. *Educational Research and Evaluation*, *8*(1), 87-108.

Lezotte, L. (1993). *Correlates of effective schools*. Maryland Educators Conference, Baltimore, MD.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1-15.312). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html adresinden edinilmiştir.

McMahon, S. D., Parnes, A. L., Keys, C. B., & Viola, J. J. (2008). School belonging among low income urban youth with disabilities: Testing a theoretical model. *Psychology in the Schools*, *45*(5), 387-401.

Mcmillian, J. H. (2015). *Sınıf içi değerlendirme* (Çev: Arı, A.). Konya: Eğitim.

MEB. (2016). *PISA 2015 ulusal rapor.* http://pisa.meb.gov.tr/wp-content/uploads/2016/12/PISA2015_Ulusal_Rapor1.pdf adresinden erişildi.

Mohammadpour, E. (2012). Factors accounting for mathematics achievement of Singaporean eighth-graders. *The Asia-Pacific Education Researcher*, *21*(3), 507-518.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide*. (7th ed.). Los Angeles, CA: Muthén ve Muthén.

Nichols, S. L. (2008). An exploration of students' belongingness beliefs in one middle school. *The journal of Experimental Education*, *76*(2), 145-169.

Oral, I. ve McGivney, E. (2013). *Türkiye'de matematik ve fen bilimleri alanlarında öğrenci performansı ve başarının belirleyicileri TIMSS 2011 analizi.* İstanbul: Eğitim Reformu Girişimi Raporu. http://erg.sabanciuniv.edu/sites/erg.sabanciuniv.edu/files/ERG%20TIMSS%202011%20Analiz%20Raporu-03.09.2013.pdf. adresinden edinilmiştir.

Organisation for Economic Co-Operation and Development [OECD]. (2004). *Learning for tomorrow's world – first results from PISA 2003.* Retrieved from: https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/34002216.pdf.

Ölçüoğlu, R. ve Çetin, S. (2016). TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *7*(1), 202-220. doi: 10.21031/epod.34424

Ölçüoğlu, R. (2015). *TIMSS 2011 Türkiye sekizinci sınıf matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi* (Yüksek lisans tezi, Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi adresinden edinilmiştir.

Özdemir, S., Sezgin, F., Şirin, H., Karip, E. ve Erkan, S. (2010). İlköğretim okulu öğrencilerinin okul iklimine ilişkin algılarını yordayan değişkenlerin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *38*(38), 213-224.

Özer, Y. ve Anıl, D. (2011). Öğrencilerin fen ve matematik başarılarını etkileyen faktörlerin yapısal eşitlik modeli ile incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *41*, 313-324.

Özgen, K. ve Bindak, R. (2011). Lise öğrencilerinin matematik okuryazarlığına yönelik öz-yeterlik inançlarının belirlenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*(2), 1073-1089.

Özüdoğru, M. (2013). *Dokuzuncu sınıf öğrencilerinin matematik başarılarının yordanması* (Yüksek lisans tezi. Ege Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitim Bilimleri Anabilim Dalı, Izmir) https://tez.yok.gov.tr/UlusalTezMerkezi adresinden edinilmiştir.

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *The Journal of Experimental Education*, *65*(3), 213-228.

Pajares, F. (2008). Motivational role of self-efficacy beliefs in self-regulated learning. In D. H. Schunk, & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory and research and applications* (1st ed., 111-140). New York: Lawrence Erlbaum Associates.

Papanastasiou, C. (2002). Internal and external factors affecting achievement in mathematics. *Studies in Educational Evaluation*, *26*, 1–7.

Peker, M. ve Mirasyedioğlu, Ş. (2003). Lise 2.sınıf öğrencilerinin matematik dersine yönelik tutumları ve başarıları arasındaki ilişki. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *14*(14), 157-166.

Sarı, M. (2012). Sense of school belonging among elemantary school students. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 42*(1), 1-11.

Sarı, M. (2013). Lise öğrencilerinde okula aidiyet duygusu. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, *13*(1), 147-160.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review*, *20*, 463–467. doi: 10.1007/s10648-008-9086-3

Strøm, I. F., Thoresen, S., Wentzel-Larsen, T., & Dyb, G. (2013). Violence, bullying and academic achievement: A study of 15-year-old adolescents and their school environment. *Child Abuse Negl, 37*(4), 243-251.

Şimşek, A. (2009). *Öğretim tasarımı*. Ankara: Nobel.

Şimşek, H. ve Katıtaş, S. (2014). İlköğretim ikinci kademe öğrencilerinde okula yabancılaşmanın çeşitli değişkenler açısından incelenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, *15*(1). 81-99.

Şişman, M., Acat, M. B., Aypay, A. ve Karadağ E. (2011). *TIMSS 2007 ulusal matematik ve fen raporu. 8. sınıflar.* Ankara MEB: EARGED Yayınları. http://yegitek.meb.gov.tr/dosyalar/dokumanlar/uluslararasi/timss_2007_ulusal_raporu.rar adresinden edinilmiştir.

Tuncer, M. ve Yılmaz, Ö. (2016). Ortaokul öğrencilerinin matematik dersine yönelik tutum ve kaygılarına ilişkin görüşlerinin değerlendirilmesi. *Kahramanmaraş Sütçü İmam Üniversitesi Sosyal Bilimler Dergisi*, *13*(2), 47-64.

Turgut, M. F. ve Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Usta, H. G. (2014). *PISA 2003 ve PISA 2012 matematik okuryazarlığı üzerine uluslararası bir karşılaştırma: Türkiye ve Finlandiya* (Doktora tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi adresinden edinilmiştir.

Uygun, N. ve Işık Tertemiz, N. (2014). Matematik dersinde probleme dayalı öğrenmenin öğrencilerin derse ilişkin tutum, başarı ve kalıcılık düzeylerine etkisi. *Eğitim ve Bilim, 39*(174), 75-90. doi: 10.15390/EB.2014.1975

Uzun, S., Bütüner, S. Ö., & Yiğit, N. (2010). A comparison of the results of TIMSS 1999-2007: The most successful five countries-Turkey sample. *Elementary Education Online*, *9*(3), 1174-1188.

Yaman, E., Eroğlu, Y., Bayraktar, B. ve Çolak, T. S. (2010). Öğrencilerin güdülenme düzeyinde etkili bir faktör: Okul zorbalığı. *Uluslararası Hakemli Sosyal Bilimler E-Dergisi, 20*, 1-17.

Yavuz, H., Demirtaşlı, R., Yalçın, S., ve İlgün Dibek, M. (2017). Türk öğrencilerin TIMSS 2007 ve 2011 matematik başarısında öğrenci ve öğretmen özelliklerinin etkileri. *Eğitim ve Bilim, 42*(189), 27-47. doi: 10.15390/EB.2017.6885

Yayan, B., & Berberoglu, G. (2004). A re-analysis of the TIMSS 1999 mathematics assessment data of the Turkish students. *Studies in Educational Evaluation*, *30*(1), 87-104.

Yıldırım, H. H., Yıldırım, S., Yetişir, M. İ. ve Ceylan, E. (2013). *PISA 2012 ulusal ön raporu*. *Millî Eğitim Bakanlığı Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü*, Ankara. http://pisa.meb.gov.tr/wp-content/uploads/2013/12/pisa2012-ulusal-on-raporu.pdf adresinden edinilmiştir.

Yılmaz, H. R. ve Bindak, R. (2016). Ortaokul öğrencilerinde matematik başarısının matematik kaygısı, sınav kaygısı ve bazı demografik değişkenlerle ilişkisinin incelenmesi. *Muğla Sıtkı Koçman Üniversitesi Eğitim Fakültesi Dergisi*, *3*(2), 30-42.

Yücel, C. ve Karadağ, E. (2016). *TIMSS 2015 Türkiye: Patinajdaki eğitim*. Eskişehir: Eskişehir Osmangazi Üniversitesi Eğitim Fakültesi. http://www.egitim.ogu.edu.tr/files/1Z5_TIMSS_2015.pdf adresinden edinilmiştir.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*(3), 265-310.

Wilkins, J. L. M., & Ma, X. (2003). Modeling change in student attitude toward and beliefs about mathematics. *The Journal of Educational Research*, *97*(1), 52-63.

Wilson, K., & Narayan, A. (2016). Relationships among individual task self-efficacy, self-regulated learning strategy use and academic performance in a computer-supported collaborative learning environment. *Educational Psychology*, *36*(2), 236-253. doi: 10.1080/01443410.2014.926312

Wood, R., & Bandura, A. (1989). Social cognitive theory of organizational management. *Academy of Management Review*, *14*(3), 361-384.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An owerview. *Theory Into Practice, 41*(2), 64-70.

EXTENDED ABSTRACT

Introduction

International studies like TIMSS give valuable information to researchers not only on mathematics and science achievement level of students but also about the relationship between student attitude, teacher and school characteristics and achievement. TIMSS (Trends in International Mathematics and Science Study) results over years indicated that Turkish students got scores below international average and Turkish educational system is not producing effective outcomes. According to recent TIMSS 2015 mathematics results, Turkish students had a ranking of 36 among 4th graders and 24 among 8th graders. There is a need to evaluate predictors of mathematics achievement to make a plan to increase achievement level of Turkish students. With these incentives, in the study it is aimed to investigate the student, teacher and school factors predicting Mathematics achievement of Turkish 8th grade students in TIMSS 2015. By determining significant and reliable factors in the prediction, the results of the study are expected to contribute stakeholders of education. The research question of which student, teacher and school characteristics are effective in predicting mathematics achievement of Turkish 8th grade students guided the study.

Method

The data of the study was obtained from student and teacher questionnaires and mathematics cognitive test scores of TIMSS 2015. In TIMSS, stratified two-stage cluster sample design was used. For the first stage, schools were sampled with probabilities proportional to their number of students. For the second stage, generally a classroom of a selected school was sampled (LaRoche, Joncas ve Foy, 2016). As a result of this sample selection method, 6079 students (48% girl and 52% boy) and 220 teachers (47% female and 53% male) attended TIMSS 2015 from Turkey. In the data analysis, multilevel regression analysis was used in which dependent variables were plausible mathematics scores and independent variables were student, teacher and school scale scores. Multilevel regression analysis was conducted by MPLUS 7.4 to identify significant predictors of mathematics achievement. MPLUS is chosen as it is capable of handling sampling characteristics of TIMSS using reported plausible scores. The explained variances of mathematics achievement accounted by student and teacher level variables were reported.

Results and Discussion

Descriptive statistics showed that Turkish students had less educational resources at home than international average. Also, students in Turkey reported that they had less than average feeling of belonging to their schools. Bullying happened rarely than average in Turkey. Additionally, students had positive view on understanding their teachers, interested in tasks given by teachers and they think that teachers use variety of interesting things in the classroom. Although Turkish students had higher level of positive attitude towards mathematics, they reported that they had less self-confidence. Teachers in Turkey reported that the emphasis given to success in schools was less than international average. Similarly, teachers reported that schools' order and safety level was less than average. When problems related to school resources were evaluated, teachers reported that Turkish schools had fewer problems. Teachers reported that they had less job satisfaction than their international colleagues. Teachers in Turkey stated that they had more problems originated from students like lack of prior knowledge and skills, lack of interest to courses, etc.

According to multilevel regression results, 34% percent of student-level variance was explained by student-level variables. It was found that self-confidence level of students was the most important predictor of mathematics achievement among student-level variables. A positive relationship was found between self-confidence and mathematics achievement. Additionally, educational resources at home variable was also among the important predictors of mathematics achievement. The students who had more educational resources at home got higher mathematics scores. The other significant variables in predicting mathematics achievement were attitude, bullying, belonging to school and engaging teaching in mathematics. Teacher and school factors explained 29% of variance at between school variations. Among these variables, school emphasis on academic success and teaching limited by student needs were two significant variables that could predict mathematics achievement of students. Students who attended schools that give more emphasis to achievement were more successful on mathematics. Students whose teachers reported more problems due to students got lower mathematics score. Other variables showed no relationship with mathematics achievement.

It is important to note that self-confidence of the students and educational resources at home were two major variables in the prediction and both of these two variables were reported to be below average in Turkish students. These two variables were considered as key to increase student success. Especially, among alterable variables student confidence could be enhanced by adequate educational strategies. Similarly, emphasis given by schools was found as a significant and important predictor of mathematics achievement which was also less than average. Schools that cooperate with parents and their students is expected to be more successful.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);266-287

Monte Carlo Simulation Studies in Item Response Theory with the R Programming Language

R Programlama Dili ile Madde Tepki Kuramında Monte Carlo Simülasyon Çalışmaları

Okan BULUT* Önder SÜNBÜL**

Abstract
Monte Carlo simulation studies play an important role in operational and academic research in educational measurement and psychometrics. Item response theory (IRT) is a psychometric area in which researchers and practitioners often use Monte Carlo simulations to address various research questions. Over the past decade, R has been one of the most widely used programming languages in Monte Carlo studies. R is a free, open-source programming language for statistical computing and data visualization. Many user-created packages in R allow researchers to conduct various IRT analyses (e.g., item parameter estimation, ability estimation, and differential item functioning) and expand these analyses to comprehensive simulation scenarios where the researchers can investigate their specific research questions. This study aims to introduce R and demonstrate the design and implementation of Monte Carlo simulation studies using the R programming language. Three IRT-related Monte Carlo simulation studies are presented. Each simulation study involved a Monte Carlo simulation function based on the R programming language. The design and execution of the R commands is explained in the context of each simulation study.

Key Words: Psychometrics, measurement, IRT, simulation, R.

Öz
Eğitimde ölçme ve psikometri alanlarında yapılan akademik ve uygulamaya dönük araştırmalarda Monte Carlo simülasyon çalışmaları önemli bir rol oynamaktadır. Psikometrik çalışmalarda araştırmacıların Monte Carlo simülasyonlarına sıklıkla başvurduğu temel konulardan birisi Madde Tepki Kuramı'dır (MTK). Geçtiğimiz son on yılda MTK ile ilgili yapılan simülasyon çalışmalarında R'ın sıklıkla kullanıldığı görülmektedir. R istatiksel hesaplama ve görsel üretme için kullanılan ücretsiz ve açık kaynak bir programlama dilidir. R kullanıcıları tarafından üretilen birçok paket program ile madde parametrelerini kestirme, madde yanlılık analizleri gibi birçok MTK temelli analiz yapılabilmektedir. Bu çalışma, R programına dair giriş niteliğinde bilgiler vermek ve R programlama dili MTK temelli Monte Carlo simülasyon çalışmalarının nasıl yapılabileceğini göstermeyi amaçlamaktadır. R programlama dilini örneklerle açıklamak için üç farklı Monte Carlo simülasyon çalışması gösterilmektedir. Her bir çalışmada, simülasyon içerisindeki R komutları ve fonksiyonları MTK kapsamında açıklanmaktadır.

Anahtar Kelimeler: Psikometri, ölçme, MTK, simülasyon, R.

INTRODUCTION

Monte Carlo simulation studies are the key elements of operational and academic research in educational measurement and psychometrics. Both academic researchers and psychometricians often choose to simulate data instead of collecting empirical data because (a) it is impractical and costly to collect the empirical data while manipulating several conditions (e.g., sample size, test length, and test characteristics); (b) it is not possible to investigate the real impact of the study conditions without knowing the true characteristics of the items and examinees (e.g., item parameters, examinee ability

* Assistant Professor, University of Alberta, Faculty of Education, Edmonton, Alberta – Canada, e-mail:bulut@ualberta.ca, ORCID ID: orcid.org/0000-0001-5853-1267
** Assistant Professor, Mersin University, Faculty of Education, Mersin – Turkey, e-mail: ondersunbul@mersin.edu.tr, ORCID ID: orcid.org/0000-0002-1775-1404

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 3, Sonbahar 201, 266-287.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 3, Autumn 2017, 266-287.
Received: 12. 04.2017
DOI: 10.21031/epod.305821
Accepted: 12. 07.2017

distributions); and (c) the empirical data are often incomplete, which may affect the outcomes of the study, especially when the amount of missing data is large and the pattern of missingness is not random (Brown, 2006; Feinberg & Rubright, 2016; Robitzsch & Rupp, 2009; Sinharay, Stern, & Russell, 2001). Furthermore, when conducting psychometric studies, it is impossible to eliminate the effects of potential confounding variables related to examinees (e.g., gender, attitudes, and motivation) and test items (e.g., content, linguistic complexity, and cognitive complexity). The growing number of research articles, books, and technical reports as well as unpublished resources (e.g., conference presentations) involving simulations also depict the importance of Monte Carlo simulations in the field of educational measurement.

Item response theory (IRT) is one of the most popular research areas in educational measurement and psychometrics. Both researchers and practitioners often use Monte Carlo simulation studies to investigate a wide range of research questions in the context of IRT (Feinberg & Rubright, 2016). Monte Carlo simulation studies are often used for evaluating how validly IRT-based methods can be applied to empirical data sets with different kinds of measurement problems (Harwell, Stone, Hsu, & Kirisci, 1996). To be able to conduct Monte Carlo simulation studies in IRT, there is a large number of psychometric software packages (e.g., IRTPRO [Cai, Thissen, & du Toit, 2011], flexMIRT [Cai, 2013], BMIRT [Yao, 2003], and Mplus [Muthén & Muthén, 1998-2015]) and programming languages (e.g., C++, Java, Python, and Fortran) available to the researchers. For some researchers, the psychometric software packages can be more suitable when conducting simulation studies in IRT because most of these packages often provide built-in functions to simulate and analyze the data. However, such psychometric software packages are often not free, only capable of particular types of IRT analyses, and generally slow when running large, computation-intensive simulations. For other researchers, the programming languages (e.g., C++ and Java) can be more tempting due the speed and flexibility, although these programming languages require intermediate to advanced programming skills to design and implement Monte Carlo simulation studies. Therefore, researchers often prefer general statistical packages – such as SAS (SAS Institute Inc., 2014), Stata (StataCorp, 2015), and R (R Core Team, 2017), which are not only more flexible and faster than the psychometric software packages but also require relatively less knowledge of programming. Among these statistical packages, R has been particularly popular because it is free, flexible, and capable of various statistical analyses and data visualizations.

Despite the increasing use of the R programming language for conducting various statistical and psychometric analyses, many researchers are still unfamiliar with the capabilities of R for conducting Monte Carlo simulation studies. As Hallgren (2013) pointed out, the use of simulation studies should be available to researchers with a broad range of research expertise and technical skills. Researchers should be familiar with how to address research questions that simulations can answer best (Feinberg & Rubright, 2016). Given the growing demand for IRT-related research in the field of educational measurement, this study aims to demonstrate how to use R for the design and implementation of Monte Carlo simulation studies in IRT, specifically for individuals with minimal experience in running simulation studies in R. The purposes of this study are threefold. First, we introduce readers the packages and functions in R for simulating response data and analyzing the simulated data using various IRT models. Second, we summarize the principles of Monte Carlo simulations and recommend some guidelines for conducting Monte Carlo simulation studies. Third, we illustrate the logic and procedures involved in conducting IRT-related Monte Carlo simulation studies in R with three examples – including the R codes for simulating item response data, analyzing the simulated data, and summarizing the analysis results. The examples will target three different uses of Monte Carlo simulation studies in IRT, including item parameter recovery, evaluating the accuracy of a method for detecting differential item functioning (DIF), and investigating the unidimensionality assumption. Each simulation study uses different criteria to evaluate the simulation results (e.g., accuracy, power, and Type I error rate). For the sake of simplicity and conciseness, the readers of this study are assumed knowledgeable about (1) the basics of the R programming language and (2) the fundamentals of IRT. The readers who are not familiar with the Monte Carlo simulation studies in IRT are referred to Harwell et al. (1996) and Feinberg and Rubright (2016) for a comprehensive review. In addition, the readers are referred to the R user manuals (https://cran.r-project.org/doc/manuals/) for a detailed introduction to the R programming language.

Some Functions to Simulate Data in R

One of the greatest advantages of R is the ability to generate variables and data sets using various probability distributions (e.g., standard normal (Gaussian) distribution, uniform distribution, and the Bernoulli distribution). This section will provide a brief summary of the probability distributions in R that are commonly used when simulating data for the IRT simulation studies. The names of the functions for generating data in R typically begin with “d” (density function), “p” (cumulative probability function), “q” (quantile function), or “r” (random sample function). The latter part of the function represents the type of the distribution. For example, the `rnorm` function generates random data with a normal distribution, while the `runif` function generates random data with a uniform distribution. The common distributions for continuous and categorical data include `exp` for the exponential distribution, `norm` for the normal distribution, `unif` for the uniform distribution, `binom` for the binomial distribution, `beta` for the beta distribution, `lnorm` for the log-normal distribution, `logis` for the logistic distribution, and `geom` for the geometric distribution.

When generating random samples using the R functions mentioned above, the randomization of the generated values is systematically controlled based on the random number generator (RNG). The RNG algorithm assigns a particular integer to each random sample. In the R programming language, the integer associated with random samples is called “seed”. The user can select a particular seed when generating a random sample and then use the same seed again whenever the same random sample needs to be obtained. The `set.seed` function can be used for selecting a particular seed in R (e.g., `set.seed(1111)`, where 1111 is the seed). The `set.seed` function plays an important role in simulations studies because it allows the researcher to create a reproducible simulation.

R Packages for Estimating IRT Models

R has many user-created psychometric packages that allow researchers and practitioners to conduct statistical analysis using psychometric models and methods. The CRAN website has a directory of the R packages categorized by topic, which is called “Task View”. One of these task views, the Psychometric Task View, is specifically dedicated to psychometric methods (see https://cran.r-project.org/web/views/Psychometrics.html), such as IRT, classical test theory, factor analysis, and structural equation modeling. A lot of the packages in the Psychometrics Task View focus on the estimation of IRT models, such as unidimensional and multidimensional IRT models, nonparametric IRT modeling, differential item functioning, and computerized adaptive testing (see Ünlü and Yanagida [2011] for a review of the CRAN Psychometrics Task View). Rusch, Mair, and Hatzinger (2013) also provided a detailed summary of the R packages for conducting IRT analysis. The primary R packages for estimating item parameters and person abilities in IRT include mirt (Chalmers, 2012), eRm (Mair, Hatzinger, & Maier, 2016), irtoys (Partchev, 2016), and ltm (Rizopoulos, 2006). There are also specific packages for a particular IRT analysis – such as lordif (Choi, Gibbons, & Crane, 2016) and difR (Magis, Beland, Tuerlinckx, & De Boeck, 2010) for differential item functioning; catR (Magis & Raiche, 2012) and mirtCAT (Chalmers, 2016) for computerize adaptive testing; and equate (Albano, 2016) for test equating. In addition to these packages, we encourage the readers to browse through the Psychometric Task View for other types of IRT methods available in R.

Guidelines for Conducting Monte Carlo Simulation Studies

Monte Carlo simulation studies can be used for investigating a wide range of research questions, such as evaluating the accuracy of existing statistical models under unfavorable conditions (e.g., small sample and non-normality), answering a novel statistical question, or understanding the empirical distribution of a particular statistic through bootstrapping (Feinberg & Rubright, 2016; Hallgren, 2013; Harwell et al., 1996). A Monte Carlo simulation study typically consists of the following steps:

1. The researcher determines a set of simulation factors expected to influence the operation of a particular statistical procedure. The simulation factors can be either fully crossed or partially crossed. If the simulation factors are fully crossed, then a data set needs to be generated for every possible combination of the simulation factors. If, however, they are partially crossed, only some simulation factors are assumed to be interacting with each other.
2. A series of assumptions are made about the nature of data to be generated (e.g., types of variables and probability distributions underlying the selected variables). These assumptions are crucial to the authenticity of the simulation study because the quality of the simulation outcomes depends on the extent to which the selected assumptions are realistic.
3. Multiple data sets are generated based on the simulation factors and the assumptions about the nature of the data. The process of generating multiple data sets is often called replication. Monte Carlo simulation studies often involve multiple replications (a) to acquire the sampling distribution of parameter estimates, (b) to reduce the chance of obtaining implausible results from a single data set, and (c) to have the option to resample the true parameters based on the assumption made in Step 2.
4. Statistical analyses are performed on the simulated data sets and the parameter estimates of interest from these analyses are recorded. The parameter estimates can be p-values, coefficients, or a particular element of the statistical model.
5. Finally, the estimated parameters are evaluated based on a criterion or a set of criteria – such as Type I error, power (or hit rate), correlation, bias, and root mean squared error (RMSE). The researcher can report the findings of the simulation study in different ways (e.g., a narrative format, tables, or graphics). The researcher should determine how the simulation results will be communicated to the target audience based on the size of the simulation study (i.e., the number of simulation factors), the complexity of the simulation design (e.g., several fully-crossed simulation factors or a simpler design with one or two factors), and the type of the reporting outlet (e.g., technical reports, journal articles, or presentations).

It should be noted that the five steps summarized above could be slightly different for each simulation study, depending on the research questions that need to be addressed.

Principles of Monte Carlo Simulation Studies

Apart from the guidelines summarized above, there are also three principles that the researchers need to consider when designing and conducting a Monte Carlo simulation study. These principles are authenticity, feasibility, and reproducibility.

The *authenticity* of a Monte Carlo simulation study refers to the degree to which the simulation study reflects the real conditions. For example, assume that a researcher wants to investigate the impact of test length on ability estimates obtained from a particular IRT model. The researcher selects 30, 60, 90, and 300 items as the hypothesized values for the test length factor. Because a 300-item test is quite unlikely to occur in real life, the researcher should probably consider eliminating this option from the simulation study. The authenticity of a Monte Carlo simulation study is also related to the necessity of the simulation factors. Continuing with the same example, the researcher might consider sample size as a potential factor for the simulation study on the recovery of ability estimates; but sample size is known to have no effect on the estimation of ability (or latent trait) when item parameters are already known (e.g., Bulut, 2013; Bulut, Davison, & Rodriguez, 2017). Therefore, the researcher would not need to include sample size as a simulation factor in the study.

The *feasibility* of a Monte Carlo study refers to the balance between the goals of the simulation study and the scope of the simulation study. The combination of many simulation factors and a high number of replications may often lead to a highly complex simulation study that is hard to complete and summarize within a reasonable period. Therefore, the researcher should determine which simulation factors are essential and how many replications can be accomplished based on the scope of the simulation study. For example, assume that a researcher plans to use test length (10, 20, 40, or 60 items), sample size (100, 250, 500, or 1000 examinees), ability distribution (normal, positively skewed, or negatively skewed), and inter-dimensional correlation (*r* = .2, *r* = .5, or *r* = .7) as the simulation factors. If the simulation factors were fully crossed, then there would be 4 x 4 x 3 x 3 = 144 cells in the simulation design. If the researcher conducted 10,000 replications for each cell, the entire simulation process would result in 1,440,000 unique data sets that need to be analyzed and summarized. Depending on the complexity of the statistical analysis, this simulation study might take several weeks (or possibly months) to complete, even with the parallel computing feature available in R and other statistical software programs.

The *reproducibility* of a Monte Carlo simulation study refers to the likelihood that the researcher who conducted the simulation study can replicate the same findings at a later time, or that other researchers who have access to the simulation parameters can replicate (or at least approximate) the findings. To ensure reproducibility, the researcher should specify the seed before generating data and store a record of the selected seeds in the simulation study. The researcher can either use the seeds to replicate the findings later on or give them to other researchers interested in replicating their findings (Feinberg & Rubright, 2016). However, it should be noted that even using the same seeds might not guarantee that identical simulation results will be obtained because the mechanism of the random number generator can differ from one computer to another, or across the different versions of the same software program.

METHOD

The following sections of this study will demonstrate three Monte Carlo simulation studies about item parameter recovery, differential item functioning, dimensionality in IRT-based assessment forms. Each study focuses on a set of research questions in the context of IRT and aims to address the research questions through a Monte Carlo simulation study. Each study consists of three steps: data generation, statistical analysis, and summarizing the simulation results. The implementation of these steps will be demonstrated using R. The readers are strongly encouraged to run the examples in their own computers by copying and pasting the provided R codes into the R console. It should be noted that most R packages are regularly updated by their creators and/or maintainers, and thus the R functions presented in this study are subject to change in the future. Therefore, we recommend the readers to check out the R packages used in this study before using the Monte Carlo simulation functions. For this study, we use the latest version of Microsoft R Open (version 3.4.0). The readers are strongly encouraged to use this particular version of Microsoft R Open to ensure the reproducibility of the Monte Carlo studies presented in the following sections.

Simulation Study 1: Item Parameter Recovery in IRT

In this study, we aim to investigate to what extent the accuracy of estimated item parameters in the unidimensional three-parameter logistic (3PL) IRT model depends on the number of examinees who respond to the items (sample size) and the number of items (test length). In addition, we want to find out which item parameter (item difficulty, item discrimination, and guessing) is the most robust against changes in sample size and test length. To address these research questions, we design a small-scale Monte Carlo simulation study in which sample size and test length are the two simulation factors. The simulation study will be based on a fully crossed design with three sample sizes (500, 1000, or 2000 examinees) and three test lengths (10, 20, or 40 items), resulting in 3 x 3 = 9 cells in total. For each cell, 100 replications will be conducted with unique item parameters and person abilities in each replication. For the evaluation of the recovery of true item parameters, we use bias and RMSE:

$$Bias = \frac{\sum_{i=1}^{K}\left(\hat{X}_i - X_i\right)}{K}, \text{and} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{K}\left(\hat{X}_i - X_i\right)^2}{K}}, \qquad (2)$$

where K is the total test length, $\hat{X}_i$ is the estimated item parameter for item i ($i = 1, 2, \ldots, K$), and X_i is the true item parameter for item i. The average bias and RMSE values over 100 replications will be reported for each of the nine simulation cells.

Data generation

The item response data will be simulated using the 3PL model. The mathematical formulation of the 3PL model can be shown as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \qquad (3)$$

where $P_i(\theta)$ is the probability of an examinee with the ability of θ responding to item *i* correctly, b_i is the item difficulty parameter for item *i*, a_i is the item discrimination parameter for item *i*, c_i is the lower asymptote (also known as the pseudo-guessing parameter), *e* is the base of the natural logarithm approximated at 2.178, and *D* is a constant of 1.7 to transform the logistic IRT scale into the normal ogive scale (Camilli, 1994; Crocker & Algina, 1986).

Following the suggestions from previous studies regarding data simulation for the 3PL model (e.g., Harwell & Baker, 1991; Feinberg & Rubright, 2016; Mislevy & Stocking, 1989; Mooney, 1997), the item difficulty parameters are drawn from a normal distribution, $b \sim N(0, 1)$; the item discrimination parameters are drawn from a log-normal distribution, $a \sim lnN(0.3, 0.2)$; and the lower asymptote parameters are drawn from a beta distribution, $c \sim Beta(20, 90)$. Furthermore, the ability parameters are drawn from a standard normal distribution, $\theta \sim N(0, 1)$. For simulating dichotomous item responses and estimating the item parameters based on the 3PL model, we will use the `mirt` package (Chalmers, 2012) in R. To install and activate the `mirt` package, the following commands should be first run:

```
install.packages("mirt")
library("mirt")
```

Then, we define a simulation function called `itemrecovery`, which generates item parameters, simulates dichotomous item response data using the generated item parameters, estimates the item parameters of the 3PL model for the simulated data, and finally computes bias and RMSE values for each set of estimated item parameters:

```
itemrecovery <- function(nitem, sample.size, seed) {

  #Set the seed and generate the parameters
  set.seed(seed)
  a <- as.matrix(round(rlnorm(nitem, meanlog = 0.3, sdlog = 0.2),3), ncol=1)
  b <- as.matrix(round(rnorm(nitem, mean = 0, sd = 1),3), ncol=1)
  c <- as.matrix(round(rbeta(nitem, shape1 = 20, shape2 = 90),3), ncol=1)
  ability <- as.matrix(round(rnorm(sample.size, mean = 0, sd = 1),3), ncol=1)

  #Simulate response data and estimate item parameters
  dat <- simdata(a = a, d = b, N = sample.size, itemtype = 'dich',
                 guess = c, Theta = ability)
  model3PL <- mirt(data=dat, 1, itemtype='3PL', SE=TRUE, verbose=FALSE)

  #Extract estimated item parameters and compute bias and RMSE
  parameters <- as.data.frame(coef(model3PL, simplify=TRUE)$items)
  bias.a <- round(mean(parameters[,1]-a), 3)
  bias.b <- round(mean(parameters[,2]-b), 3)
  bias.c <- round(mean(parameters[,3]-c), 3)
  rmse.a <- round(sqrt(mean((parameters[,1]-a)^2)), 3)
  rmse.b <- round(sqrt(mean((parameters[,2]-b)^2)), 3)
  rmse.c <- round(sqrt(mean((parameters[,3]-c)^2)), 3)

  #Combine the results in a single data set
  result <- data.frame(sample.size=sample.size, nitem=nitem,
                       bias.a=bias.a, bias.b=bias.b, bias.c=bias.c,
                       rmse.a=rmse.a, rmse.b=rmse.b, rmse.c=rmse.c)
  return(result)
}
```

In the `itemrecovery` function, there are three input values that need to be specified by the researcher: `nitem` as the number of items (i.e., test length), `sample.size` as the number of examinees (i.e., sample size), and `seed` as the seed for the random number generator. The `itemrecovery` function begins with setting the seed for the random values that are going to be generated, which will ensure reproducibility of the simulated data. Next, the item parameters are randomly generated based on the distribution characteristics explained earlier using the `rlnorm`, `rnorm`, and `rbeta` functions. Each set of the generated parameters (called a, b, and c) is saved as a matrix with a single column. The `simdata` function from the `mirt` package simulates dichotomous item responses according to the 3PL model using the generated item parameters. More details about the `simdata` function can be obtained by running the `?simdata` command in the R console. Then, we estimate item parameters using the `mirt` function, extract the estimated item parameters from the model using the `coef` function, and save the parameters in a data frame called `parameters`. More details about the estimation process in the `mirt` function can be obtained by running the `?mirt` command in the R console. At the end of the function, we compute the bias and RMSE values for each item parameter, save the values into a data set called `result`, and return the `result` data set as the outcome of the simulation. To enable the `itemrecovery` function, we can either copy and paste the entire function into the R console and hit the “enter” button, or select all the entire function in the R script file, right-click on the selected lines, and choose “Run line or selection” to execute the commands in the R console.

The next step is to conduct the simulation study using the `itemrecovery` function. First, we randomly generate 100 integers to be used as the random seeds in the study. We sample random integers ranging from 0 to 1,000,000 using the `sample.int` function and store the generated values in a data set called `myseed`. We export the seeds into a text file called `"simulation seeds.txt"`. This file will be saved in the current working directory designated by the user. To save the document in a specific folder, a complete folder path should be provided, such as `“C:/Users/username/Desktop/simulation  seeds.txt”`. Note that when defining a folder path, forward slash (/) should be used instead of a backslash (\). Next, we define an empty data set (i.e., `result`) that will store the simulation results out of 100 replications. The final step of the simulation study is to run the simulation within a loop and save the results into the `result` data set. `for (i in 1:length(myseed)){ }` creates a loop to run a procedure 100 times (i.e., the same length of `myseed`). In this study, we want to run the `itemrecovery` function 100 times using a different seed from `myseed` for each replication. Once all the iterations are complete, the `result` data set will consist of one hundred rows (one row per iteration). At the end, we use the `colMeans` function to find the average bias and RMSE values across 100 replications. Because we use the `round(colMeans(result),3)`, all of the values will be rounded off to three decimal digits.

```
#Generate 100 random integers
myseed <- sample.int(n = 1000000, size = 100)
write.csv(myseed, "simulation seeds.txt", row.names = FALSE)
```

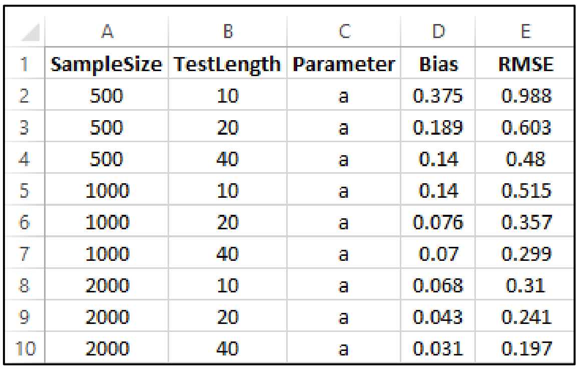
```
#Define an empty data frame to store the simulation results
result <- data.frame(sample.size=0, nitem=0, bias.a=0, bias.b=0, bias.c=0,
                     rmse.a=0, rmse.b=0, rmse.c=0)

#Run the loop and return the results across 100 iterations
for (i in 1:length(myseed)) {
  result[i,] <- itemrecovery(nitem = 20, sample.size = 1000, seed = myseed[i])
}
round(colMeans(result), 3)
```

For each of the nine cells planned for this simulation study, we set the `nitem` and `sample.size` values accordingly and re-run the R script presented above. When the simulations for all of the cells are complete (i.e., the R script has been run nine times in total), we report the results shown in Table 1. The simulation results show that the lower asymptote (i.e., guessing) parameter had the smallest bias and RMSE, whereas the item discrimination parameter indicated the largest bias and RMSE. As sample size and test length increased, bias and RMSE decreased for all of the item parameters.

Table 1. Simulation Results from the Item Parameter Recovery Study

Sample Size	Test Length	Bias a	Bias b	Bias c	RMSE a	RMSE b	RMSE c
500	10	0.375	-0.250	-0.004	0.988	0.931	0.160
	20	0.189	-0.139	-0.003	0.603	0.658	0.144
	40	0.140	-0.109	-0.001	0.480	0.555	0.132
1000	10	0.140	-0.091	-0.012	0.515	0.589	0.134
	20	0.076	-0.046	-0.004	0.357	0.428	0.122
	40	0.070	-0.063	0.001	0.299	0.384	0.109
2000	10	0.068	-0.033	-0.010	0.310	0.376	0.112
	20	0.043	-0.031	-0.006	0.241	0.302	0.092
	40	0.031	-0.020	-0.004	0.197	0.264	0.085

In addition to Table 1, we can also present the simulation results graphically using the `lattice` package (Sarkar, 2008) in R. First, we manually enter the simulation results in Table 1 into an empty Excel spreadsheet using a long format and save the spreadsheet with a .csv extension using the "Save As" option under the "File" menu. The saved file is called "simulation results.csv". Figure 1 shows a screenshot of the "simulation.csv" file.



	A	B	C	D	E
1	SampleSize	TestLength	Parameter	Bias	RMSE
2	500	10	a	0.375	0.988
3	500	20	a	0.189	0.603
4	500	40	a	0.14	0.48
5	1000	10	a	0.14	0.515
6	1000	20	a	0.076	0.357
7	1000	40	a	0.07	0.299
8	2000	10	a	0.068	0.31
9	2000	20	a	0.043	0.241
10	2000	40	a	0.031	0.197

Figure 1. A Screenshot of the First Nine Rows of the "simulation results.csv" File

After we read the data set "simulation results.csv" in R, we define the two variables (`SampleSize` and `TestLength`) as categorical variables by using the `as.factor` function. Next, we install the `lattice` package and then activate it using the `library` command. Finally, we use the `xyplot` function in the `lattice` package (Sarkar, 2008) to create an interaction plot. This plot will demonstrate the relationship between bias, RMSE, sample size, and test length for each item parameter. In the `xyplot` function, we first select the variable for the y axis (either bias or RMSE) and the variable for the x axis (test length), the variable that defines multiple panels (sample size), and the group variable (parameters). In addition, `xlab` defines the label for the x axis, `type = "a"` indicates an interaction plot, the elements in `auto.key` define the position of the legend, whether or not data points should be shown, whether or not lines should be shown, and the number of columns for the legend, and the elements in `par.settings` defines the colours (`lty`) and thickness (`lwd`) of the lines in the plot. The details of the `xyplot` function can be obtained by running the `?xyplot` command in the R console. Figures 2 and 3 show the interaction plots for bias and RMSE, respectively.

```r
#Reading in "parameter recovery.csv"
result <- read.csv("parameter recovery.csv", header = TRUE)
result$SampleSize <- as.factor(result$SampleSize)
result$TestLength <- as.factor(result$TestLength)

#Create interaction plots using the lattice package
install.packages("lattice")
library("lattice")
xyplot(Bias ~ TestLength | SampleSize,result,group=Parameter,xlab="Test Length",
type = "a",auto.key=list(corner=c(1,0.9),points=FALSE,lines=TRUE,columns=1),
par.settings=simpleTheme(lty=1:3,lwd=2))

xyplot(RMSE ~ TestLength | SampleSize,result,group=Parameter,xlab="Test Length",
type = "a",auto.key=list(corner=c(1,0.9),points=FALSE,lines=TRUE,columns=1),
par.settings = simpleTheme(lty=1:3,lwd=2))
```

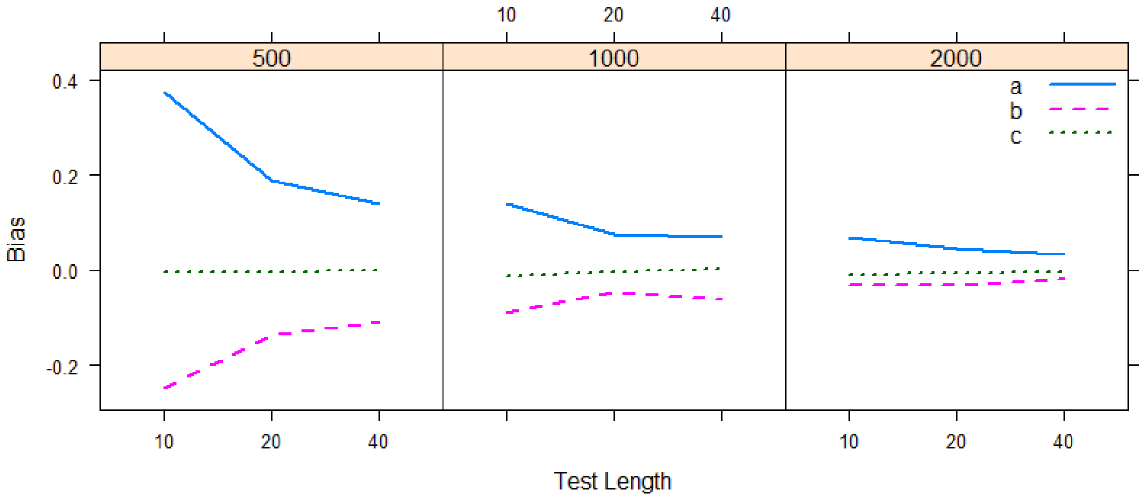


Figure 2. The Interaction Plot for Bias, Test Length, and Sample Size

Summary

This simulation study investigated the effects of sample size and test length on the recovery of item parameters from the 3PL model. To demonstrate how to evaluate the accuracy of estimated item parameters in R, this study included two simulation factors (test length and sample size) with 100 replications. The results suggested that both test length and sample size are negatively associated with the accuracy of item discrimination and item difficulty parameters. As sample size and test length increased, both bias and RMSE decreased. Unlike item discrimination and item difficulty parameters, the lower asymptote (guessing) parameter was slightly affected by the simulation factors. Future studies can focus on item parameter recovery by expanding the simulation factors of the current study (e.g., smaller or larger sample sizes), using different IRT models, such as Graded Response Model (Samejima, 1969), adding other simulation factors (e.g., ability distribution, extreme guessing parameters for the 3PL model, and non-simple structure in multidimensionality).

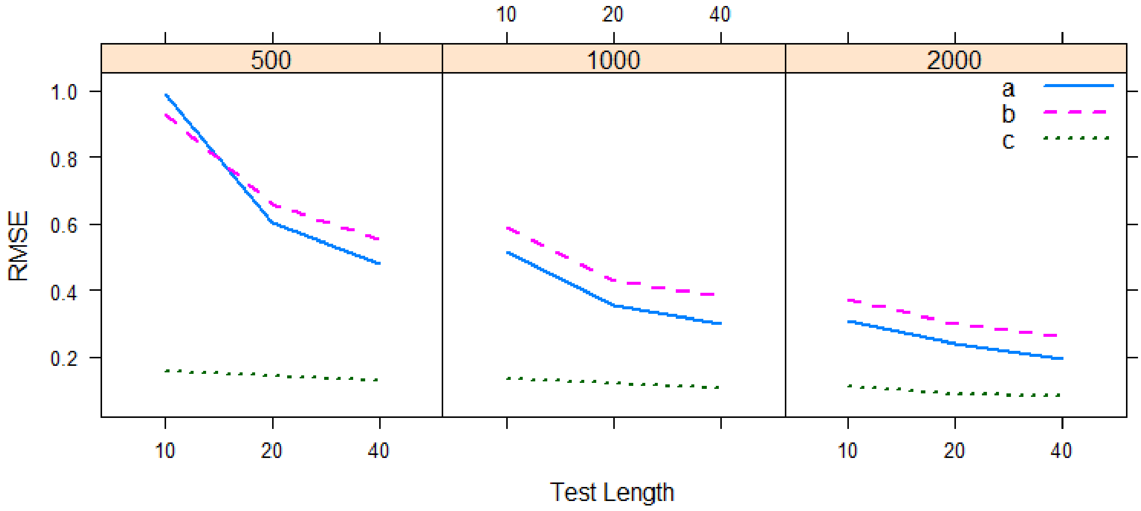


Figure 3. The Interaction Plot for RMSE, Test Length, and Sample Size

Simulation Study 2: Detecting Differential Item Functioning in Multidimensional IRT

The second simulation study aims to investigate the detection of differential item functioning (DIF) in the context of multidimensional IRT models. In educational testing, DIF occurs when the probability of responding to a dichotomous item correctly varies between focal and reference groups (e.g., male and female students), after controlling for examinees' ability levels. If the item is polytomous, then the probabilities of obtaining no credit, a partial credit (e.g., 1 point in a two-point item), or a full credit (e.g., 2 points in a two-point item) are expected to differ between the focal and reference groups, after controlling for examinees' ability levels. There are two types of DIF in test items: uniform and nonuniform DIF. If the focal group consistently underperforms or outperforms the reference group, then the item is flagged for having uniform DIF. If, however, the direction of bias changes between the focal and reference groups along the ability continuum, the item is flagged for having nonuniform DIF (Lee, Bulut, & Suh, 2016).

There are many methods in the literature to detect uniform and nonuniform DIF in the context of unidimensional IRT models. These methods include the Mantel-Haenszel method (Mantel & Haenszel, 1959), simultaneous item bias test (SIBTEST; Shealy & Stout, 1993), Raju's differential functioning of items and tests (DFIT; Raju, van der Linden, & Fleer, 1995); and the multiple indicators multiple causes (MIMIC) model (Finch, 2005; Woods & Grimm, 2011). However, when the definition of DIF is extended to a multidimensional assessment that simultaneously measures two or more abilities, there are only a few DIF methods in the literature, such as multidimensional MIMIC-interaction model (Lee et al., 2016), IRT likelihood ratio test (Suh & Cho, 2014), and multidimensional SIBTEST (MULTISIB; Stout, Li, Nandakumar, & Bolt, 1997).

In this Monte Carlo simulation study, we use the IRT likelihood ratio test described by Suh and Cho (2014) for detecting uniform and nonuniform DIF in the context of multidimensional Graded Response Model (MGRM). The mathematical formulation of MGRM for a polytomous item with $K+1$ response categories on an M-dimensional test becomes:

$$P_{ik}^{*}(\boldsymbol{\theta}) = \frac{1}{1+e^{[-D\sum_{m=1}^{M}\mathbf{a_{im}}(\theta_m - b_{ik})]}}, \tag{4}$$

where $P^*_{ik}(\boldsymbol{\theta})$ is the probability of selecting the response option k $(k = 1, \dots, K)$ in item i for an examinee with the ability vector of $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]$, b_{ik} is the boundary parameter of the k^{th} category of item i, a_{im} is the item discrimination parameter for item i on dimension m $(m = 1, \dots, M)$, and D is a constant of 1.7 to transform the logistic IRT scale into the normal ogive scale.

In this study, we assume a two-dimensional test with a simple structure. The test consists of 30 items, where the first set of 15 items is loaded on the first dimension and the second set of 15 items is loaded on the second dimension. To simulate polytomous item responses, we use item characteristics similar to those from Jiang, Wang, and Weiss's (2016) recent study on MGRM. Item discrimination parameters are randomly drawn from a uniform distribution, $a \sim U(1.1, 2.8)$; the first category boundary parameter is randomly drawn from a uniform distribution, $b_1 \sim U(0.67, 2)$ and the other two category boundary parameters are created by subtracting a value randomly drawn from a uniform distribution, $b_2 = b_1 - U(0.67, 1.34)$ and $b_3 = b_2 - U(0.67, 1.34)$. In addition, the ability parameters are drawn from a multivariate normal distribution, $\boldsymbol{\theta} \sim MVN(0, \boldsymbol{\Sigma})$ where is $\boldsymbol{\Sigma}$ the variance-covariance matrix of the abilities.

Three simulation factors are manipulated in this study: sample size, DIF magnitude, and inter-dimensional correlation. Sample size is manipulated for the reference (R) and focal (F) groups as R1000/F200, R1500/F500, and R1000/F1000. DIF magnitude is manipulated as 0, 0.3, or 0.6 logit difference for uniform and nonuniform DIF. DIF magnitude is added to the category boundary parameters for uniform DIF and to the item discrimination parameters for nonuniform DIF. We assume that the focal group is at a disadvantage due to uniform or nonuniform DIF. Inter-dimensional correlation refers to the correlation between the two ability dimensions. Inter-dimensional correlation is manipulated as ρ=0, ρ=.3, or ρ=.5. Based on the value of the inter-dimensional correlation, the off-diagonal elements of the variance-covariance matrix ($\boldsymbol{\Sigma}$) are replaced with ρ, while the diagonal elements remain as "1".

For simulating polytomous item responses, estimating the item parameters based on MGRM, and running the IRT likelihood ratio tests, we will use the `MASS` package (Venables & Ripley, 2002) and the `mirt` package (Chalmers, 2012) in R (R Core Team, 2017). In addition, we will use the `doParallel` package (Revolution Analytics & Weston, 2015) to benefit from the parallel computing to further speed up the estimation process. To install and then activate these packages, the following commands should be first run in the R console:

```r
install.packages("doParallel")
library("doParallel")
library("mirt")
library("MASS")
```

Next, we define a function called `detectDIF`, which generates item parameters and simulates polytomous item responses based on MGRM, estimates the item parameters using the simulated data, runs IRT likelihood ratio tests to detect uniform and nonuniform DIF on a particular set of items, and computes the true positive rates (i.e., power) and false positive rates (i.e., Type I error) as the evaluation criteria.

The `detectDIF` function requires four input values: `sample.size` defines the size of reference and focal groups (e.g., `sample.size = c(1000, 200)` for the reference group of 1000 examinees and the focal group of 200 examinees); `DIF.size` defines the magnitude of uniform DIF and nonuniform DIF (e.g., `DIF.size = c(0.3, 0)` for 0.3 difference in the category threshold parameters as uniform DIF and `DIF.size = c(0, 0.3)` for 0.3 difference in the discrimination parameters as nonuniform DIF); `cor` specifies the correlation between the two ability dimensions (e.g., `cor = 0.5` for a correlation of 0.5 between the two abilities); and `seed` is the user-defined seed for the data generation process. Based on the selected input values, the function generates a 30-item, polytomously-scored test in which items 1, 7, 15, 16, 23, and 30 are tested for uniform and nonuniform DIF. These items are particularly selected because they represent a combination of low-, medium-, and high-difficulty as well as low-, medium-, and high-discrimination parameters.

The IRT likelihood ratio test examines the likelihood difference between two nested IRT models based on a chi-square test with degrees of freedom equal to the difference between the numbers of estimated item parameters between the two IRT models (see Suh and Cho [2014] for more details on this procedure). To investigate uniform DIF with the IRT likelihood ratio test, we first estimate `mod0` that assumes that all items except for items 1, 7, 15, 16, 23, 30 are invariant between the focal and reference group. Next, we estimate `mod1` that constrains the category boundary parameters b1, b2, and b3 to be equal between the focal and reference group for each of the six DIF items and tests whether there is a significant change in the model likelihood due to the equality constraints. Significant likelihood changes from `mod0` to `mod1` indicate that the items being tested exhibit uniform DIF. Nonuniform DIF is examined by comparing `mod0` against `mod1` and `mod2`, which constrain the item discrimination parameters a1 and a2 to be equal between the focal and reference groups, respectively. Significant likelihood changes between these models indicate that the items being tested exhibit nonuniform DIF.

The `detectDIF` function returns a data frame in which the aforementioned six items and the p-values from the IRT likelihood ratio test for each item are listed. When one of the input values for `DIF.size` is larger than zero, the function returns the p-values for detecting DIF correctly (i.e., power) for the six items listed above. If, however, `DIF.size = c(0, 0)`, then the function returns the p-values for detecting DIF falsely (i.e., Type I error) for the items. In this study, we assume that DIF occurs either in the category boundary parameters or in the item discrimination parameters. Therefore, one of the values in `DIF.size` will always be zero when running the analysis for power.

```r
detectDIF <- function(sample.size, DIF.size, cor, seed) {

  require("mirt")
  require("MASS")
  set.seed(seed)

  #Define multidimensional abilities for reference and focal groups
  theta.ref <- mvrnorm(n = sample.size[1], rep(0, 2), matrix(c(1,cor,cor,1),2,2))
  theta.foc <- mvrnorm(n = sample.size[2], rep(0, 2), matrix(c(1,cor,cor,1),2,2))

  #Generate item parameters for reference and focal groups
  a1 <- c(runif(n = 15, min = 1.1, max = 2.8), rep(0,15))
  a2 <- c(rep(0,15), runif(n = 15, min = 1.1, max = 2.8))
  a.ref <- as.matrix(cbind(a1, a2), ncol = 2)
  b1 <- runif(n = 30, min = 0.67, max = 2)
  b2 <- b1 - runif(n = 30, min = 0.67, max = 1.34)
  b3 <- b2 - runif(n = 30, min = 0.67, max = 1.34)
  b.ref <- as.matrix(cbind(b1, b2, b3), ncol = 3)

  #Uniform and nonuniform DIF for items 1, 7, 15, 16, 23, and 30
  b.foc <- b.ref
  b.foc[c(1,7,15,16,23,30),] <- b.foc[c(1,7,15,16,23,30),]+DIF.size[1]
  a.foc <- a.ref
  a.foc[c(1,7,15),1] <- a.foc[c(1,7,15),1]+DIF.size[2]
  a.foc[c(16,23,30),2] <- a.foc[c(16,23,30),2]+DIF.size[2]

  #Generate item responses according to MGRM
  ref <- simdata(a = a.ref, d = b.ref, itemtype = 'graded', Theta = theta.ref)
  foc <- simdata(a = a.foc, d = b.foc, itemtype = 'graded', Theta = theta.foc)
  dat <- rbind(ref, foc)
  #Define the group variable (0=reference; 1=focal) and test DIF using mirt
  group <- c(rep("0", sample.size[1]), rep("1", sample.size[2]))
  itemnames <- colnames(dat)
  model <- 'f1 = 1-15
            f2 = 16-30
            COV = F1*F2'
  model.mgrm <- mirt.model(model)
  #Test uniform DIF
  if(DIF.size[1]>0 & DIF.size[2]==0) {
    mod0 <- multipleGroup(data = dat, model = model.mgrm, group = group,
                          invariance = c(itemnames[-c(1,7,15,16,23,30)],
                                         'free_means', 'free_var'), verbose = FALSE)
    mod1 <- DIF(mod0, c('d1','d2','d3'), items2test = c(1,7,15,16,23,30))

    result <- data.frame(items=c(1,7,15,16,23,30),
                         DIF=c(mod1[[1]][2,8], mod1[[2]][2,8], mod1[[3]][2,8],
                               mod1[[4]][2,8], mod1[[5]][2,8], mod1[[6]][2,8]))
  } else

    #Test nonuniform DIF
    if(DIF.size[1]==0 & DIF.size[2]>0) {
      mod0 <- multipleGroup(data = dat, model = model.mgrm, group = group,
                            invariance = c(itemnames[-c(1,7,15,16,23,30)],
                                           'free_means', 'free_var'), verbose = FALSE)
      mod1 <- DIF(mod0, c('a1'), items2test = c(1,7,15))
      mod2 <- DIF(mod0, c('a2'), items2test = c(16,23,30))
      result <- data.frame(items=c(1,7,15,16,23,30),
                           DIF=c(mod1[[1]][2,8], mod1[[2]][2,8], mod1[[3]][2,8],
                                 mod2[[1]][2,8], mod2[[2]][2,8], mod2[[3]][2,8]))
    } else

      #Test type I error
      if(DIF.size[1]==0 & DIF.size[2]==0) {
        mod0 <- multipleGroup(data = dat, model = model.mgrm, group = group,
                              invariance = c(itemnames[-c(1,7,15,16,23,30)],
                                             'free_means', 'free_var'), verbose = FALSE)
        mod1 <- DIF(mod0, c('a1','d1','d2','d3'), items2test = c(1,7,15))
        mod2 <- DIF(mod0, c('a2','d1','d2','d3'), items2test = c(16,23,30))
        result <- data.frame(items=c(1,7,15,16,23,30),
                             DIF=c(mod1[[1]][2,8], mod1[[2]][2,8], mod1[[3]][2,8],
                                   mod2[[1]][2,8], mod2[[2]][2,8], mod2[[3]][2,8]))
      }
  return(result)
}
```

For this study, we use 30 replications, as in Jiang et al.'s (2016) simulation study with MGRM. Despite using only 30 replications, this simulation study is more complex compared to the first simulation study presented earlier because in addition to estimating item parameters from a two-dimensional MGRM, we conduct a series of IRT likelihood ratio tests to examine uniform and nonuniform DIF across six items (items 1, 7, 15, 16, 23, and 30). To increase the speed of the entire simulation process, we use the `doParallel` package. First, we generate a set of 30 random seeds ranging from 0 to 10000 and save the generated seeds in a data set called "`myseed`". Next, using a computer with a multi-core processor, we allocate multiple cores for our simulation study. To check the number of processors in a computer, the researcher can first run `detectCores()`. To assign a particular number of cores, `registerDoParallel()` should be used. For example, to allocate 8 cores for the simulation study, `registerDoParallel(8)` should be used. Once this command is executed, the simulation process can be completed using 8 cores rather than a single core, which is the default setting in R. Although it is possible to use all available cores in a computer, this could be problematic because using all available cores would slow down the operation of the computer significantly, especially when performing other tasks to in addition the simulation study. The parallel computing is particularly useful when an iterative computing process – such as a simulation study – is implemented. Because the current simulation study requires 30 replications, estimating multiple replications simultaneously is expected to reduce the duration of simulation significantly (e.g., with 8 cores, it would be theoretically 8 times faster than a regular estimation with a single core).

```
myseed <- sample.int(n=10000, size = 30)
detectCores()
registerDoParallel(8)
```

Once the parallel computing process is set up, the final step of this simulation study is to run the simulation study by changing the input values in the detectDIF function. The foreach function from the doParallel package is used for creating a loop of 30 replications. After each replication, the results will be combined in a single data set called result. We use the ifelse function to create a binary variable for the items in the result data set that have a p-value less than .05 (i.e., significant DIF based on the IRT likelihood ratio test). The mean function will return the proportion of the items that have indicated significant DIF (i.e., average power or Type I error depending on the condition). The following R commands demonstrate an example for running a particular combination of the simulation factors. For each replication, a random seed from myseed is selected. Then, the detectDIF function is executed using sample sizes of 1000 and 200 for the reference and focal groups, the DIF magnitude of 0.3 for uniform DIF, and a correlation of ρ=.3 between the two ability dimensions. After 30 replications are complete, the average proportion of significant IRT likelihood ratio tests across six items is reported with two decimal points (using the round function).

```
result <- foreach (i = 1:30, .combine=rbind) %dopar% {
  detectDIF(sample.size=c(1000, 200), DIF.size=c(0.3, 0), cor=0.3, seed=myseed[i])
}
round(mean(ifelse(result$DIF < 0.05, 1, 0)),2)
```

Table 2. Power and Type I Error Rates for Detecting DIF in MGRM

Sample Size	Correlation	Power for Uniform DIF		Power for Nonuniform DIF		Type I Error
		0.3	0.6	0.3	0.6	
R1000/F200	0	.23	.80	.21	.51	.09
	.3	.28	.84	.23	.46	.07
	.5	.31	.82	.22	.49	.06
R1500/F500	0	.53	1	.39	.83	.07
	.3	.56	.98	.40	.82	.04
	.5	.55	.98	.44	.81	.04
R1000/F1000	0	.60	.98	.44	.66	.06
	.3	.56	.98	.37	.68	.03
	.5	.64	.97	.43	.73	.06

Table 2 shows a summary of the findings across all simulation factors. The results show that sample size and DIF magnitude are positively associated with the power of the IRT likelihood ratio test when detecting uniform and nonuniform DIF in MGRM. As sample size and DIF magnitude increased, power rates of the IRT likelihood ratio for detecting both uniform and nonuniform DIF test also increased. Unlike sample size and DIF magnitude, the effect of inter-dimensional correlation does not seem to be consistent regarding power rates. When DIF magnitude is large (i.e., 0.6), the correlation between the two dimensions has no effect on power rates for uniform DIF. However, the correlation between the dimensions affects power rates slightly for nonuniform DIF. Overall, the IRT likelihood ratio test appears to detect uniform DIF more precisely than nonuniform DIF. Type I error rates appear to be reasonable, although small sample size condition (R1000/F200) seems to have higher Type I error rates than the other two sample size conditions. As for the power rates, the effect of inter-dimensional correlation is not consistent regarding Type I error rates. The results in Table 2 can also be summarized with a scatterplot to demonstrate the relationships between the simulation factors more visually (see the graphical example in Simulation Study 1 for more details.).

Summary

This simulation study investigated the effects of sample size, DIF magnitude, and inter-dimensional correlation in detecting uniform and nonuniform DIF for a multidimensional polytomous IRT model (i.e., MGRM). Power rates in detecting DIF correctly and Type I error rates in detecting DIF falsely are used as the evaluation criteria. For the demonstration purposes, we only used 30 replications but the number of replications could be easily increased with the help of parallel computing. This would result in more reliable simulation results. Future studies can expand the simulation factors of the current study to have a more comprehensive analysis. For example, different values for DIF magnitude, sample size, and inter-dimensional correlations can be used. Furthermore, new simulation factors can be included. For example, three or higher dimensional structures can be used to see the impact of the number of dimensions, which would also allow examining the impact of varying inter-dimensional correlations. In addition, instead of a simple structure, a complex test structure with items associated with multiple dimensions can be assumed.

Simulation Study 3: Investigating Unidimensionality

The third simulation study aims to investigate test dimensionality. Unidimensionality is essential for the test theories like as Classical Test Theory or Item Response Theory. Therefore, the investigation of test dimensionality is very important. The unidimensionality assumption requires that there is a single latent trait underlying a set of test items. As Hambleton, Swaminathan, and Rogers (1991) pointed out, the unidimensionality assumption may not hold for most measurement instruments in education, psychology, and other social sciences due to complex cognitive and non-cognitive factors, such as motivation, anxiety, and ability to work quickly. Therefore, we can expect that at least one minor extra factor confounds unidimensionality. However, finding a major component or a factor underlying the data is adequate to meet the unidimensionality assumption.

Monte Carlo studies can be very convenient for investigating the factors affecting unidimensionality under various conditions. In the following Monte Carlo study, we aim to examine the impact of sample size, the number of items associated with a secondary (nuisance) dimension, and inter-dimensional correlations on the detection of unidimensionality. Two-dimensional response data with a simple structure are generated. While most items are assumed to be associated with the first ability dimension, the number of items (10, 20, or 30 items) associated with a secondary dimension is manipulated as a simulation factor. Second, the correlation between the two ability dimensions (ρ=.3, ρ=.6, or ρ=.9) is manipulated. As the third simulation factor, sample size (500, 1000, or 3000 examinees) is modified because sample size is considered an important factor for the accuracy of dimensionality analyses. The three simulation factors are fully crossed, resulting in 3 x 3 x 3 = 27 cells in total. One hundred replications are conducted for each cell.

A multidimensional two-parameter logistic IRT model (M2PL) is used for data generation. The M2PL model can be written as follows:

$$P_i(\boldsymbol{\theta}) = \frac{\exp(\sum_{m=1}^{M} a_{im}\theta_{\mathrm{m}} + d_i)}{1 + \exp(\sum_{m=1}^{M} a_{im}\theta_{\mathrm{m}} + d_i)}, \tag{5}$$

where $P_i(\boldsymbol{\theta})$ is the probability of responding to item *i* correctly for an examinee with the ability vector of $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]$, a_{im} is the discrimination parameter of item *i* related to ability dimension *m* (*m* = 1, 2, …, *M*), and d_i is the difficulty parameter of item *i*. In this study, the item discrimination parameters are randomly drawn from a uniform distribution, $a{\sim}U(1.1, 2.8)$, the item difficulty parameters are randomly drawn from a uniform distribution, $d{\sim}U(0.67, 2.00)$, and the ability values are obtained from a multivariate normal distribution, $\boldsymbol{\theta}{\sim}MVN(0, \boldsymbol{\Sigma})$ where is $\boldsymbol{\Sigma}$ the variance-covariance matrix of the abilities. Each generated data set is analyzed with NOHARM explanatory factor analysis (McDonald, 1997), which is an effective method for finding the number of underlying dimensions in item response data (Finch & Habing, 2005). NOHARM is implemented with one factor restriction using the sirt package (Robitzsch, 2017). The average root mean square error of approximation (RMSEA) is used as the evaluation criterion. RMSEA values smaller than .05 are usually considered a close fit, whereas RMSEA values equal or greater than .10 are considered a poor fit (Browne & Cudeck, 1993; Hu & Bentler, 1999).

For this study, we define a function called `detectDIM`, which draws item difficulty and item discrimination parameters according the distributions explained earlier, simulates two-dimensional response data with a simple structure based on the M2PL model, fits an exploratory factor model with a one-factor restriction (i.e., unidimensional model) to the simulated data, and extracts the RMSEA value as the evaluation criterion. Before using the `detectDIM` function, the following packages must be installed and activated:

```
install.packages("sirt")
library("sirt")
library("mirt")
library("MASS")
```

The `detectDIM` function requires five input values: `sample.size` for the number of examinees, `testLength1` for the number of items with a non-zero loading on the first ability dimension, `testLength2` for the number of items with a non-zero loading on the second ability dimension, `cor` for the correlation between the two ability dimensions, and `seed` for setting the random number generator. The `detectDIM` function is shown below:

```
detectDIM <- function(sample.size, testLength1, testLength2, cor, seed) {

    require("mirt")
    require("MASS")
    require("sirt")
    set.seed(seed)

    # Generate item discrimination and difficulty parameters
    a1 <- c(runif(n = testLength1, min = 1.1, max = 2.8), rep(0, testLength2))
    a2 <- c(rep(0, testLength1), runif(n = testLength2, min = 1.1, max = 2.8))
    disc.matrix <- as.matrix(cbind(a1, a2), ncol = 2)
    difficulty <- runif(n = (testLength1 + testLength2), min = 0.67, max = 2)

    # Specify inter-dimensional correlations
    sigma <- matrix(c(1, cor, cor, 1), 2, 2)

    # Simulate data
    dataset <- simdata(disc.matrix, difficulty, sample.size, itemtype = 'dich',
                       sigma = sigma)

    # Analyze the simulated data by fitting a unidimensional model
    noharmOneFactorSolution <- noharm.sirt(dat = dataset , dimensions = 1)

    # Summarize the results
    result <- data.frame(sample.size = sample.size , testLength1 = testLength1,
                         testLength2 = testLength2, cor = cor,
                         RMSEA = noharmOneFactorSolution$rmsea)

    return(result)
}
```

To start the simulation study, we first generate 100 random seeds ranging from 0 to 1,000,000 and save the seeds into a file called “simulation seeds.csv”. The numbers stored in this file will be useful if we want to replicate the findings of this study in the future. Next, we create an empty data frame called “results.csv” and save this file into the current directory. This is a comma-separated-values file, which can be opened with any text editor or Microsoft Excel. This file will store all average RMSEA values across the 27 simulation cells. For each cell, 100 replications will be temporarily stored in a data set called “`result`” and the average RMSEA values from this data set will be stored in the results.csv file. Unlike in the first two simulation studies presented earlier, the input values in this simulation study are entered into nested loops so that these input values do not have to be modified manually. For example, `for (ss in 1:length(sample.size))` creates a loop with three sample size values (see `sample.size <- c(500, 1000, 3000)`). These values will be used iteratively in the `detectDIM` function via `sample.size = sample.size[ss]`. In addition to the loop for sample size, there are three other loops for test length for the second dimension, correlations between the ability dimensions, and the replications based on the random seeds, respectively. The simulation process will stop automatically after 100 replications are completed for each of the 27 cells, and it will write the average RMSEA values into the results.csv file.

```
myseed <- sample.int(n = 1000000, size = 100)
write.csv(myseed, "simulation seeds.csv", row.names = FALSE)
write.table(matrix(c("Sample Size", "Test Length 1", "Test Length 2",
                    "Correlation","Mean RMSEA"), 1, 5), "results.csv", sep = ",",
            col.names = FALSE, row.names = FALSE)
# Create an empty data frame to save the results
result <- data.frame(sample.size = 0, testLength1 = 0, testLength2 = 0,
                     cor = 0, RMSEA = 0)

# Run all the input values through loops
sample.size <- c(500, 1000, 3000)
test.length2 <- c(10, 20, 30)
correlation <- c(0.3, 0.6, 0.9)
for (ss in 1:length(sample.size)) {
  for (tl in 1:length(test.length2)) {
    for (k in 1:length(correlation)) {
      for (i in 1:length(myseed)) {
        result[i, ] <-detectDIM(sample.size = sample.size[ss],testLength1 = 30,
                            testLength2 = test.length2[tl], cor = correlation[k],
                            seed = myseed[i])
      }
      meanRMSEAs <- round(colMeans(result), 3)
      write.table(matrix(meanRMSEAs, 1, 5), "results.csv", sep = ",",
                  col.names = FALSE, row.names = FALSE, append = TRUE)
    }
  }
}
```

After all of the iterations (27 cells x 100 replications = 2700 iterations in total) are completed, we read “results.csv” and call the data set “`summary`” in R. Then, we use the `summary` data set to create a graphical summary of the findings through the `dotplot` function in the `lattice` package. To consider the simulation values as labels in the graph, we use the `factor` function, which saves a numerical variable as a character variable. In addition, we use the `paste0` function to make the labels more clear. For example, instead of using 500, 1000, and 3000 as the labels, we combine these values with the text “Sample Size=”, and create the following labels: Sample Size=500, Sample Size=1000, and Sample Size=3000. With the `levels` option, it is possible to set the order of the created labels, which changes in which order the labels will appear in the graph. The `dotplot` function creates a scatterplot of the correlation between the dimensions and average RMSEA values. The vertical line “|” between `Mean.RMSEA` and `Sample.Size*Test.Length.2` allows us to create a separate scatterplot for each sample size and test length 2 combination.

```
# Read in the summary results in
summary <- read.csv("results.csv", header = TRUE)
summary$Sample.Size <- factor(paste0("Sample Size=",summary$Sample.Size),
                              levels = c("Sample Size=500", "Sample Size=1000",
                                         "Sample Size=3000"))
summary$Test.Length.2 <- factor(paste0("Test Length 2=",summary$Test.Length.2))
summary$Correlation <- factor(summary$Correlation)

library(lattice)
```

```
dotplot(Correlation ~ Mean.RMSEA | Sample.Size*Test.Length.2 , data = summary,
        pch=c(2), cex=1.5, xlab = "Average RMSEA", aspect=0.5, layout = c(3,3),
        ylab="Correlation", xlim=c(0,0.3))
```

Figure 4 shows the results of the Monte Carlo simulation study across the three simulation factors. The results indicate that test length and inter-dimensional correlations can influence the dimensional structure of item response data. When the two ability dimensions are highly correlated (i.e., ρ=.9), average RMSEA values are less than 0.05, suggesting a close fit for the one-factor model. However, as the correlation between the ability dimensions decreases and the number of items associated with the secondary dimension increases, RMSEA values become substantially larger, suggesting a poor fit for the one-factor model. Unlike the test length and correlation factors, sample size does not appear to affect the magnitude of RMSEA. In Figure 4, as the sample size increases from 500 (first column from the left) to 3000 (last column from the left), the average RMSEA values remain nearly the same, holding the other two simulation factors constant.

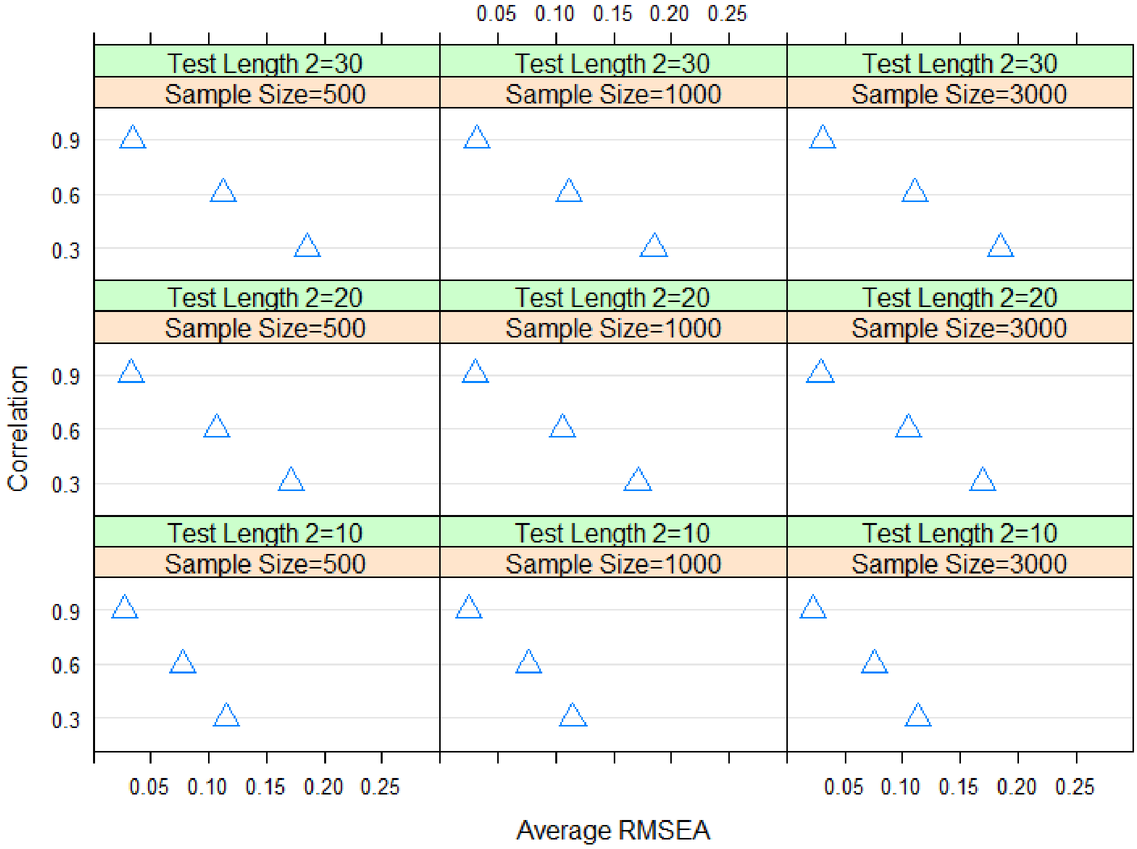


Figure 4. Average RMSEA Values across the Three Simulation Factors

Summary

This simulation study investigated the impact of a secondary dimension on the detection of unidimensionality. The simulation study involved three simulation factors: the number of items related to the secondary dimension, the correlation between the ability dimensions, and sample size. The M2PL model was used as the underlying model for data generation. A one-factor model was fit to each simulated data set using the `noharm.sirt` function in the `sirt` package and the RMSEA values were extracted as the evaluation criterion. Future studies can expand the scope of the current study. For example, different values for test length (e.g., fewer or more items) and inter-dimensional correlations (e.g., ρ=0) can be used. In addition, new simulation factors can be included to investigate different research questions. For example, the number of secondary (nuisance) dimensions can be influential on the detection of unidimensionality. This would also enable the use of varying correlations between dimensions since there would be more values to modify in the variance-covariance matrix of the abilities. Finally, instead of a simple test structure, secondary dimensions can be used for generating a complex test structure where most items are dominantly loaded on the first dimension but the items also share some variance with the secondary dimensions.

DISCUSSION and CONCLUSION

Academic researchers and practitioners often use Monte Carlo simulation studies for investigating a wide range of research questions related to IRT. During the last decade, R has become one of the most popular software programs for designing and implementing Monte Carlo simulation studies in IRT. The R programming language allows researchers to simulate various types of data, analyze the generated data based on a particular model or method of interest, and summarize the results statistically and graphically. Given the growing popularity of R among researchers and practitioners, this study provided a brief introduction to the R programming language and demonstrated the use of R for conducting Monte Carlo studies in IRT. Each simulation study presented in this study focuses on a different aspect of IRT, involves a variety of simulation factors, and uses various criteria to evaluate the outcomes of the simulations. We recommend the readers either to conduct the same Monte Carlo simulation studies or to design their own simulation studies by following the R codes provided in this study. In addition, the readers who are interested in the nuts and bolts of the R programming language are encouraged to check out many R resources available on the internet (e.g., https://journal.r-project.org/ and http://www.statmethods.net/).

REFERENCES

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.). *Testing structural equation models*. Newbury Park, CA: Sage.

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation, University of Minnesota).

Bulut, O., Davison, M. L., & Rodriguez, M. C. (2017). Estimating between-person and within-person subscore reliability with profile analysis. *Multivariate Behavioral Research, 52*(1), 86–104.

Cai, L. (2013). *flexMIRT version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Lincolnwood, IL: Scientific Software International.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in item response theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293–295.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1–39.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). lordif: Logistic ordinal regression differential item functioning using IRT [Computer software]. Available from https://CRAN.R-project.org/package=lordif.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston, Inc.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36–49.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278–295.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement, 42*, 149–169.

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology, 9*(2), 43–60.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. New York: Sage publications.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, *15*(4), 375–389.

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology, 7*(109), 1–10.

Lee, S., Bulut, O., & Suh, Y. (2016). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*. Advance online publication. DOI: 10.1177/0013164416651116.

Mair, P., Hatzinger, R., & Maier M. J. (2016). *eRm: Extended Rasch modeling* [Computer software]. Available from http://CRAN.R-project.org/package=eRm.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862.

Magis, D., & Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R Package catR. *Journal of Statistical Software, 48*(8), 1–31.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. van der Linden & R. K. Hambleton, *Handbook of modern item response theory* (pp. 257-269). New York: Springer.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57–75.

Mooney, C. Z. (1997). *Monte Carlo simulations*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-116. Thousand Oaks, CA: Sage.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide (7th Edition)*. Los Angeles, CA: Muthén & Muthén.

Partchev, I. (2016). irtoys: A collection of functions related to item response theory (IRT) [Computer software]. Available from https://CRAN.R-project.org/package=irtoys.

R Core Team (2017). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from https://cran.r-project.org/.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with application of differential item functioning. *Applied Psychological Measurement*, *19*, 353–368.

Revolution Analytics, & Weston, S. (2015). *doParallel: Foreach parallel adaptor for the 'parallel' package* [Computer software]. Available from https://CRAN.R-project.org/package=doParallel.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Robitzsch, A. (2017). *sirt: Supplementary item response theory models* [Computer software]. Available from https://CRAN.R-project.org/package=sirt.

Robitzsch, A., & Rupp, A. A. (2009). The impact of missing data on the detection of differential item functioning. *Educational Psychological Measurement, 69*, 18–34.

Rusch, T., Mair, P., & Hatzinger, R. (2013). *Psychometrics with R: A review of CRAN packages for item response theory*. Retrieved from http://epub.wu.ac.at/4010/1/resrepIRThandbook.pdf

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

SAS Institute Inc. (2014). *Version 9.4 of the SAS system for Windows*. Cary, NC: SAS Institute Inc.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DTF. *Psychometrika*, *58*, 159–194.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*, 317–329.

StataCorp. (2015). *Stata statistical software: Release 14*. College Station, TX: StataCorp LP.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB – A procedure to investigate DIF when a test is intentionally multidimensional. *Applied Psychological Measurement*, *21*, 195–213.

Suh, Y., & Cho, S. J. (2014). Chi-square difference tests for detecting functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement, 38*, 359–375.

Ünlü, A., & Yanagida, T. (2011). R you ready for R?: The CRAN psychometrics task view. *British Journal of Mathematical and Statistical Psychology, 64*(1), 182–186.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York, NY: Springer.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*, 339–361.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Computer software]. Monterey, CA: Defense Manpower Data Center. Available from http://www.bmirt.com.

UZUN ÖZET

Giriş

Son yıllarda eğitimde ölçme ve psikometri alanlarındaki çalışmalar incelendiğinde Monte Carlo simülasyon çalışmalarının oldukça büyük yer tuttuğu görülmektedir. Bu çalışmalar, özellikle teoriye yönelik olan psikometrik çalışmalar için vazgeçilmez bir rol üstlenmiş konumdadır. Bunun başlıca nedenleri Monte Carlo simülasyon çalışmalarının sağlamış olduğu avantajlardan kaynaklanmaktadır. Bu avantajlar; simülasyon koşullarını değişimleyecek şekilde görgül veri toplamanın uygulamada mümkün olmaması veya maliyetinin yüksek olması, madde ve/veya birey parametrelerinin gerçek değerlerinin görgül veri kapsamında tam olarak bilinememesinden kaynaklı oluşan sıkıntılar, görgül verinin kayıp verilerin yarattığı sorunlar, görgül verilerde karıştırıcı değişkenlerden arınık bir ortam sağlanamamasından kaynaklı olarak neden-sonuç ilişkilerinin kurulmasında yaşanan zorluklar olarak sırlanabilir. Son yıllarda yaşanan bir diğer gelişme ise R programı ve programlama dilinde oluşan gelişmelerdir. Ücretsiz ve açık kaynak kodlu olan R, başta istatistik olmak üzere, birçok alanda yaygın olarak kullanılmaya başlanmıştır. Psikometride R'ın yaygın olarak kullanılmaya başlandığı alanlar arasında yer almaktadır. R'ın bel kemiğini oluşturan CRAN web sayfasında yer alan dizinlerden biri olan "Task View", R'da yer alan paketlerin konulara göre kategorize edildiği bir alt alandır. Bu dizinde yer alan Psychometrics bağlantısı açıldığında, psikometriye yönelik olarak birçok çalışmanın yapılmış olduğu görülebilir. Bu çalışmalar kapsamında madde tepki kuramına (MTK) dair yer alan kestirimlere yönelik başlıca paketlerin, mirt (Chalmers, 2012), eRm (Mair, Hatzinger, & Maier, 2016), irtoys (Partchev, 2016), and ltm (Rizopoulos, 2006) olduğu söylenebilir. Bunun yanında bazı özelleştirilmiş MTK paketlerinden de söz etmek mümkündür: Değişen madde fonksiyonu için lordif (Choi, Gibbons, & Crane, 2016) ve difR (Magis, Beland, Tuerlinckx, & De Boeck, 2010); bilgisayar ortamında bireye uyarlanmış testler için catR (Magis & Raiche, 2012) ve mirtCAT (Chalmers, 2016); test eşitleme için equate (Albano, 2016) ve MTK için birçok açıdan destekleyici ve bağlantı kurucu nitelikte olan sirt (Robitzsch, 2017) bu paketler arasında gösterilebilir. R programlama dili kullanarak, R'da psikometriye yönelik paketler kullanarak birçok IRT tabanlı Monte Carlo çalışması yürütülmüştür ve hala da yürütülmektedir. Bu çalışmanın amacı IRT'ye dayalı veri üretimi ve incelemesini içeren simülasyon çalışmalarının R'da nasıl yapılması gerektiğine dair bilgilendirme sağlamak ve örnekler üzerinden elde edilmiş olan bulguları paylaşmaktır.

Yöntem

Bu çalışma kapsamında üç adet Monte Carlo simülasyon çalışması yürütülmüştür. Bu çalışmalardan birincisi parametre yeniden elde edimine yönelik olup, ikincisi değişen madde fonksiyonuna yönelik bir çalışmadır. Üçüncü çalışma ise çeşitli faktör yapılarındaki tek boyutluluğun incelenmesi üzerinedir.

Birinci simülasyon çalışmasında; tek boyutlu 3 parametreli lojistik modele dayalı olarak veri üretimi yapılmış ve bu üretimlerden elde edilen madde parametrelerinin orijinal madde parametrelerine olan benzeşikliği incelemiştir. İncelemeler yanlılık ve RMSE kapsamında gerçekleştirilmiştir. Çalışma kapsamında örneklem büyüklüğü (500, 1000, 2000) olacak şekilde ve test uzunluğu (10, 20, 40) olacak şekilde değişimlenmiştir. Simülasyon çerçevesinde b parametresi standart normal dağılımdan, $b \sim N(0, 1)$; ayırt edicilik parametresi olan a parametresi, log-normal dağılımdan, $a \sim lnN(0.3, 0.2)$; ve düşük asimptot parametresi olan c beta dağılımından , $c \sim Beta(20, 90)$ elde edilmiştir.

İkinci simülasyon çalışmasında; tek biçimli ve tek biçimli olmayan değişen madde fonksiyonu içeren çoklu puanlanan maddeler barındıran veri üretimi gerçekleştirilmiştir. Üretim çok boyutlu aşamalı tepki modeline dayalı olarak 30 madde için yapılmıştır. İlk 15 madde birinci faktöre ait olup ikinci 15 madde ise ikinci faktöre ait olacak şekilde basit yapı formatında üretim yapılmıştır. Tek biçimli değişen madde fonksiyonu için odak ve referans gruplarının verilerinin eşik değerleri farklılaştırılmıştır. Tek düzeyli olmayan değişen madde fonksiyonu verisi üretirken ise belirtilen grupların madde ayırt edicilik değerleri de farklılaştırılmıştır. Simülasyon kriteri olarak; örneklem büyüklüğü (R1000/O200, R1500/O500 ve R1000/F1000), Değişen madde fonksiyonu logit büyüklüğü (0, 0.3, 0.6). Olabilirlik testi kullanılarak güç ve birinci tip hata incelemesi gerçekleştirilmiştir.

Üçüncü simülasyon çalışmasında ise çok boyutlu MTK'ya dayalı olarak 2 boyutlu, basit yapı formatında, iki kategorik maddelerden oluşan üretim gerçekleştirilmiştir. İlk boyutta yer alan madde sayısı sabit tutulmakla beraber ikinci boyutta yer alan maddelerin sayısı (10, 20, 30) olacak şekilde değişimlenmiştir. Diğer simülasyon kriterleri ise örneklem büyüklüğü (500,1000, 3000) ve boyutlar arasındaki korelasyondur (0.3, 0.6, 0.9). Simülasyon için kullanılan a ve d parametreleri uniform dağılımdan elde edilirken $a \sim U(1.1, 2.8)$, $d \sim U(0.67, 2.00)$, örtük özelliklere ilişkin yetenek dağılımları tanımlanan korelasyonalar bağlı olarak, ortalamaları 0 olan çok değişkenli normal dağılımdan $\boldsymbol{\theta} \sim MVN(0, \boldsymbol{\Sigma})$ elde edilmiştir. Üretilmiş olan veriler sirt (Robitzsch, 2017) paketi kullanılarak analiz edilmiştir. Verilere, tek faktör sınırlandırması altında açımlayıcı NOHARM faktör analizi uygulanmıştır. Yapılmış olan çözümlemelerden elde edilen RMSEA değerleri boyutluluk değerlendirmesi için kullanılmıştır.

Sonuçlar ve Tartışma

Birinci simülasyon çalışması sonucunda, madde ayırt edicilik parametreleri ve madde güçlük parametrelerinin kestirim uygunluğunun, test uzunluğu ve örneklem büyüklüğü ile negatif yönde bağlantı gösterdiği gözlenmiştir. Başka bir deyişle test uzunluğu ve örneklem büyüklüğü arttıkça elde edilen yanlılık ve RMSE değerleri düşmüştür. Buna ek olarak düşük asimptot parametresi olan c parametresinin kestirim uygunluğu simülasyon kriterlerinden oldukça az etkilenmiştir.

İkinci simülasyon çalışmasının sonucunda, örneklem büyüklüğünün ve değişen madde fonksiyonu büyüklüğünün, olabilirlik testi sonuçlarının gücüyle pozitif yönde ilişki gösterdiği söylenebilir. Başka bir deyişle örneklem büyüklüğü ve değişen madde fonksiyonu büyüklüğü arttıkça olabilirlik testinin de gücü artmaktadır. Boyutlar arası korelasyon ve güç arasında tutarlı bir ilişkilendirme sağlanamamıştır. Çalışmadan elde edilen diğer bir sonuç ise tek biçimli olan değişen madde fonksiyonu için olan sonuçların güç açısından daha yüksek olmasıdır. 1. Tip Hata sonuçları için, genel olarak kabul edilebilir sınırlar etrafında olduğu, görece daha düşük örneklem büyüklüklerinde hata oranın arttığı, boyutlar arası korelasyona dayalı incelemelerin güç çalışmasındakine benzer bir şekilde tutarlı bir sonuç vermediği söylenebilir.

Üçüncü simülasyon çalışmasından elde edilen sonuçlar incelendiğinde, ikinci boyutta yer alan madde sayısının ve boyutlar arasındaki korelasyonun verinin boyutluluk yapısını etkilediği görülmüştür. Boyutlar arasındaki korelasyon arttıkça, veri için yapılan tek boyutlu çözümlemenin daha uygun olduğu görülmüştür. Bunula beraber boyutlular arasındaki korelasyon azaldığında ve ikinci boyuttaki madde sayısı arttığında verinin tek boyutluluktan uzaklaştığı gözlemlenmiştir. Son olarak örneklem büyüklüğünün, boyutluluğu diğer faktörler kadar etkilemediği görülmüştür.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);288-302

Creating Parallel Forms to Support On-Demand Testing for Undergraduate Students in Psychology

Mark J. GIERL*　　Lia DANIELS**　　Xinxin ZHANG***

Abstract
On-demand testing requires that multiple forms of an exam should be administered to students in each testing session. However, the use of multiple forms raises test security concern because of item exposure. One way to limit exposure is using parallel forms construction. Parallel forms are different versions of a test that measure the same content areas and have the same difficulty level but contain different sets of items. The purpose of this study is to describe and demonstrate how parallel forms can be created from a small item bank using the selected-response item type. We present three unique yet plausible test assembly problems. We also provide a solution for each problem using the results from a free, open-source, software add-in for Microsoft Excel called the Opensolver. Implications for test design and item development are discussed.

Key Words: Test development, automated test assembly, fairness in testing

INTRODUCTION

The principles and practices that guide the design and development of educational tests are undergoing dramatic changes. One of the major catalysts for this change stems from the application of technology in assessment that is best exemplified with the rapid development and application of computer-based testing (CBT) (Drasgow, 2016; Drasgow, Luecht, & Bennett, 2006; Luecht, 2016; Sireci & Zenisky, 2016). A computer-based test is an exam that contains digitally-formatted items that are delivered with a computer. The test administration can be conducted using a computer network or over the Internet. Computer-based tests may contain traditional selected-response (e.g., multiple choice) and constructed-response (e.g., short answer, essays) item types. They may also contain new “innovative” item types—such as drag-and-drop, reordering, and multiple select—that are only made possible with a computer administration because they require technology-based features such as interactivity or multimedia (Sireci & Zenisky, 2016). In short, CBT is dramatically changing educational assessment because the use of expanded item types combined with the growing popularity of digital media and the explosion in Internet use is creating the foundation for a new type of testing system. As a result, many educational tests that were once given in a paper format are now administered by computer as either computer-based or computer adaptive exams. Many common and well-known exams can be cited as examples including ACT Aspire, the Graduate Management Achievement Test, the Graduate Record Exam, the Test of English as a Foreign Language, the American Institute of Certified Public Accountants Uniform CPA examination, the Medical Council of Canada Qualifying Exam Part I, the National Council Licensure Examination for Registered Nurses, and the Canadian Practical Nurse Registration Examination.

Benefits and Challenges Related to Computer-Based Testing

CBT is growing in popularity not just with large-scale testing companies but also with individual instructors who recognize that paper-based testing is an exceedingly time- and resource-intensive

* Professor of Educational Psychology and Canada Research Chair, University of Alberta, Faculty of Education, Edmonton-Canada, e-mail: mark.gierl@ualberta.ca , ORCID ID: 0000-0002-2653-1761
** Associate Professor of Educational Psychology, University of Alberta, Faculty of Education, Edmonton-Canada, e-mail:lia.daniels@ualberta.ca, ORCID ID: 0000-0001-9202-2538
*** Doctoral Student, University of Alberta, Faculty of Education, Edmonton-Canada, xinxin4@ualberta.ca:, ORCID ID: 0000-0003-4926-7980

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 3, Sonbahar 2017, 288-302.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 3, Autumn 2017, 288-302.

Received:10.04.2017
Accepted:24.08.2017

DOI: 10.21031/epod.305350

process. The printing, scoring, and reporting of paper-based tests require tremendous efforts, expenses, and human interventions. Moreover, as the demand for testing continues to escalate, the cost of developing, administering, and scoring paper-based tests will also continue to increase. One solution that curtails some of these costs is to adopt a CBT system. By administering tests on computers, instructors are liberated from performing the costly and time-consuming administration processes associated with disseminating, scanning, and scoring paper-based tests.

CBT offers many benefits to students and instructors compared to more traditional paper-based testing. Computers permit testing on-demand thereby allowing students to take the exam using a flexible test administration schedule. This benefit means that students can take exams at different locations due to the flexibility of networked computer and Internet access. It also means that instructors can administer tests outside of their scheduled lecture times so that no classes are lost due to a required test administration. Selected-response item types like multiple-choice are scored immediately by the computer. As a result, instructors are not required to manually grade the exams or send the exams to an external facility for optical scoring. If permitted by the instructor, students can even be provided with feedback on their performance immediately upon the completion of the exam because all exams are scored by the computer (Daniels & Gierl, in press). Computer-based tests can be accessed with different types of devices ranging from mobile technologies such as tablets to standalone desktop computers typically found in computer labs or testing centres. Computers also support the development of “innovative” item types that allows teachers to measure more complex performances as well as a broader variety of knowledge and skills using diverse interactive item formats.

Despite these important benefits, two major challenges may prevent instructors initially from attempting to implement CBT. First, instructors may feel that their item bank is too small. A bank is a repository of test items. It contains information about content areas measured by each item as well as statistical information (e.g., difficulty level) about the performance of each item. The items in this repository are used to create forms. On-demanding testing requires that multiple versions or forms of the exam are administered to students in each testing session. Students are then randomly assigned one of the forms. Second, instructors may be concerned that on-demand CBT compromises the security of the exam by allowing students to see the items. These exposed items, in turn, could be disclosed to other students who take a different form or take the exam at a different time. One way to effectively mitigate this concern is to create parallel test forms. Parallel forms are different versions of the test that measure the same content areas and have the same difficulty level but contain a different set of test items (van der Linden, 1998, 2006). The purpose of our study is to describe and demonstrate how instructors can create and implement parallel forms construction using selected-response item types like multiple-choice items. Parallel test forms are considered to be secure exams because each form contains a different set of items thereby minimizing item exposure and maintaining test security so the test administration is fair and equitable for all students[1]. Because parallel forms measure the same content areas and produce tests with the same difficulty levels, instructors can compare students’ test results because the scores across the forms are deemed to be equivalent. We begin by providing a description of the construction process using the selected-response item type. Then, we present software that can used to easily implement the construction process using a small bank of test items.

Test Form Assembly

Producing parallel forms that yield both reliable and valid test scores is complicated because a complex combinatorial problem must be solved. Using a manual test assembly approach, the instructor would first identify the content areas measured by the test, use either the statistical item analysis results from previously administered test items or attempt to predict the expected difficulty

[1] To further enhance security, the order of the items on the parallel forms can also be randomized so students who write the same form receive the same items, but in a different order.

level of the items, and, finally, create multiple forms that each contain items that measure the same content areas and displayed the same item difficulty values. Specifying an item that measures the same content area and displays the same difficulty level from a pool of possible items is a form of constraint programming. As tests increase in specificity with the inclusion of a larger number of content areas and as the number of items increases producing a broader range of item difficulty levels, more constraints are required to produce parallel forms. In addition, multiple content specifications may be required to create truly parallel forms including the use of constraints for variables such as test length, item format, item exposure, date of creation, and source of item. The use of these additional content specifications makes the test assembly process a daunting and, potentially, impossible one to solve using a manual process because items must be selected to meet a statistical requirement while also satisfying two or more content specifications.

Fortunately, efficient computer-based procedures for automated test assembly (ATA) have been developed (see van der Linden, 1998, 2005). Using the optimization algorithms found in a range of readily available software, instructors can quickly solve complex test assembly problems in order to produce parallel test forms, even with relatively small item banks. To date, however, the widespread use of ATA techniques, particularly among university instructors, has been limited. This limitation may be attributed to the instructors lack of knowledge about ATA or it may be attributed to their lack of confidence for implementing the ATA process. The purpose of our study is to address these limitations by describing and illustrating how ATA methods can be used by instructors in psychology to create parallel test forms.

Overview of Automated Test Assembly

ATA requires the optimization of a test attribute (e.g., overall mean test score) using a unique combination of items so that a feasible solution (e.g., item combinations that meet the content specifications) is produced. The problem of selecting items to meet predefined test specifications can be approached using computerized combinatorial optimization methods (Breithaupt & Hare, 2016, Luecht, 1998; van der Linden, 1998, 2005). Optimization requires the specification of a mathematical model to describe the combinatorial problem. One approach for articulating these models is to specify a system of linear equations that defines the decision variables, an objective function, and the constraints. The system of linear equations contains both integer and string variables. Integers (i.e., discrete variables) are used to describe item attributes such as content area, test length, and item format. Strings (i.e., continuous variables) are used to describe statistical attributes such as mean test score. Once a model is defined, mixed-integer linear programming methods are used to iteratively assess every possible solution relative to the target until the optimal or best combination is identified.

Program Description with Example

Instructors can solve test assembly problems using optimization algorithms found in specialized software such as IBM ILOG CPLEX, LINGO 16.0, or Premium Solver Pro. In our study, the Opensolver (https://opensolver.org/) add-in was used because it is a freely available, open-source tool than can be run within Microsoft Excel to solve mixed integer optimization problems. Opensolver was created and is supported by Andrew Mason and Iain Dunning in the Department of Engineering Science at the University of Auckland.

Two steps are required to assemble parallel forms using Opensolver. The first step is to describe the test assembly problem as a mathematical model in an Excel spreadsheet. The second step is to specify and solve the test assembly model in the Solver Model interface. To begin, we provide a simple test assembly problem with simulated data from the selected-response item type to illustrate how an ATA problem is defined and specified using the Opensolver. In problem 1, two parallel forms containing 35 selected-response items were assembled from a hypothetical item bank containing 100 items. The bank contained the content code and the difficulty level for each item.

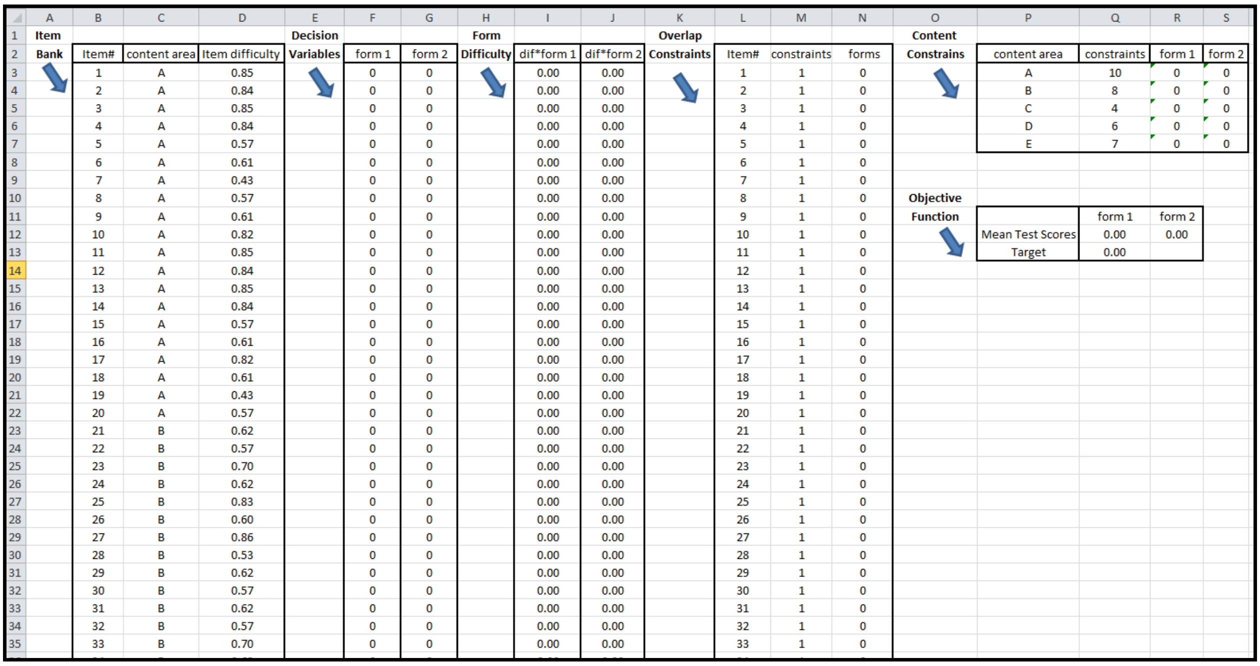
Every time an instructor administers a multiple-choice exam, statistics can be computed for each item including the difficulty level. The difficulty level is computed as the proportion of students who answered the item correctly. It ranges from 0.00 to 1.00. Both forms in problem 1 were assembled to meet the same content and statistical specifications. Five content areas were measured on each test. Each form of the exam required 35 items from five different content areas—10 items from content area A, eight items from area B, four items from area C, six items from area D, and seven items from area E. The two forms were also required to be statistically parallel meaning that the items on each form must produce the same mean test score. The requirements for this test assembly problem are summarized in Table 1. Manually selecting two sets of 35 mutually exclusive items from a bank of 100 items that produce the same mean test score across the same five content areas is not a simple task. As an alternative, OpenSolver can be used to solve this test assemble problem very quickly.

Table 1. Summary of Content and Statistical Requirements for Problem 1

Content Area	Number of Items in Bank	Difficulty Mean (SD)	Number of Items Per Form
A	20	0.70 (0.15)	10
B	20	0.65 (0.11)	8
C	20	0.65 (0.15)	4
D	20	0.74 (0.17)	6
E	20	0.72 (0.23)	7
Total	100	0.69 (0.16)	35

Step 1: Setting up the mathematical model in Excel

To begin, the content and statistical data for each item in the bank must be specified in an Excel spreadsheet, as shown in the item bank table in Figure 1 (this figure displays 33 of the 100 items). The bank contains the item number, content area, and difficulty level for each item. For example, Cell B3, C3 and D3 indicate item 1 belongs to content A and it has a difficulty level of 0.85 (very easy item). Next, the decision variables (which includes the form difficulty and item overlap tables), the content constraints, and the objective function of the ATA problem must be specified. The decision variables table is used to specify the ATA model for the problem. These cells represent the item decision variable matrix (i.e., item-by-test form). The cells contain the values that the solver algorithm begins with and then updates in order to search for an optimal solution. In our example, the cells are constrained to be binary so that a 1 means that an item was selected for the required form and a 0 means it was not selected. The starting values for these cells are 0, which means that the items were not selected. In step 2 where we specify and solve the test assembly model (shown in next section), these values will change thereby describing the solution for the ATA problem (i.e., Figure 1 and 3 are used together to describe the initial state and the final solution to our problem). For example, Cell F3=0 means that item 1 was not selected for form 1. Cell G4=0 means that item 2 was not selected for form 2. The form difficulty and item overlap tables are defined using the same logic. The starting point for form difficulty is that the difficulty level of each item begins with the value of 0.00. We also define item overlap. In problem 1, we do not want any overlap among the items across the two forms (i.e., the forms will each contain unique mutually exclusive items from the bank). We define item overlap by specifying the starting point for the constraint to be 1 (column M) and the form to be 0 (column N).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**Item**				**Decision**			**Form**			**Overlap**				**Content**				
2	**Bank**	Item#	content area	Item difficulty	**Variables**	form 1	form 2	**Difficulty**	dif*form 1	dif*form 2	**Constraints**	Item#	constraints	forms	**Constrains**	content area	constraints	form 1	form 2
3		1	A	0.85		0	0		0.00	0.00		1	1	0		A	10	0	0
4		2	A	0.84		0	0		0.00	0.00		2	1	0		B	8	0	0
5		3	A	0.85		0	0		0.00	0.00		3	1	0		C	4	0	0
6		4	A	0.84		0	0		0.00	0.00		4	1	0		D	6	0	0
7		5	A	0.57		0	0		0.00	0.00		5	1	0		E	7	0	0
8		6	A	0.61		0	0		0.00	0.00		6	1	0					
9		7	A	0.43		0	0		0.00	0.00		7	1	0					
10		8	A	0.57		0	0		0.00	0.00		8	1	0	**Objective**				
11		9	A	0.61		0	0		0.00	0.00		9	1	0	**Function**		form 1	form 2	
12		10	A	0.82		0	0		0.00	0.00		10	1	0		Mean Test Scores	0.00	0.00	
13		11	A	0.85		0	0		0.00	0.00		11	1	0		Target	0.00		
14		12	A	0.84		0	0		0.00	0.00		12	1	0					
15		13	A	0.85		0	0		0.00	0.00		13	1	0					
16		14	A	0.84		0	0		0.00	0.00		14	1	0					
17		15	A	0.57		0	0		0.00	0.00		15	1	0					
18		16	A	0.61		0	0		0.00	0.00		16	1	0					
19		17	A	0.82		0	0		0.00	0.00		17	1	0					
20		18	A	0.61		0	0		0.00	0.00		18	1	0					
21		19	A	0.43		0	0		0.00	0.00		19	1	0					
22		20	A	0.57		0	0		0.00	0.00		20	1	0					
23		21	B	0.62		0	0		0.00	0.00		21	1	0					
24		22	B	0.57		0	0		0.00	0.00		22	1	0					
25		23	B	0.70		0	0		0.00	0.00		23	1	0					
26		24	B	0.62		0	0		0.00	0.00		24	1	0					
27		25	B	0.83		0	0		0.00	0.00		25	1	0					
28		26	B	0.60		0	0		0.00	0.00		26	1	0					
29		27	B	0.86		0	0		0.00	0.00		27	1	0					
30		28	B	0.53		0	0		0.00	0.00		28	1	0					
31		29	B	0.62		0	0		0.00	0.00		29	1	0					
32		30	B	0.57		0	0		0.00	0.00		30	1	0					
33		31	B	0.62		0	0		0.00	0.00		31	1	0					
34		32	B	0.57		0	0		0.00	0.00		32	1	0					
35		33	B	0.70		0	0		0.00	0.00		33	1	0					

Figure 1. Model Specified for the ATA Problem in Problem 1.

Next, the content constraints table is used to define the content on the parallel forms. The number of items measuring each content category on each test is calculated in the columns (R, S). These cells are defined so they equal the sum of the decision variables assigned to items under each specific category. For instance, cell R3 is equal to sum of F(3:22) because these decision variables are for items measuring content A, which are items 1 to 20 in the bank. The values of the cells in these columns (R, S) must be equal to the values in column Q, as specified in the content constraints table.

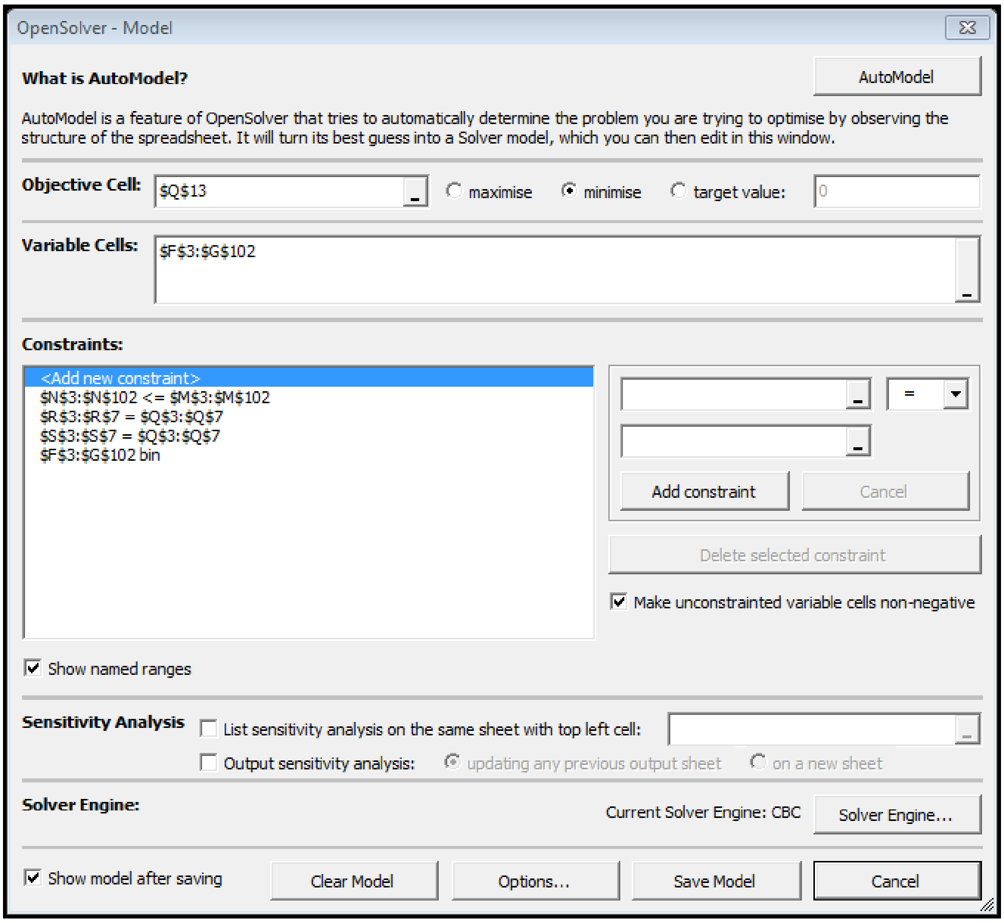
Finally, the objective function table is defined. This function specifies the statistical requirement for our problem. In order to make the two forms have the same mean test score, the objective function was formulated to minimize the absolute difference between the mean score on the two forms. Hence the difference between the forms is presented as P13 = 0.00, meaning the forms should have the same mean test score. The mean test score for each form is calculated in Cells Q12 and R12 based on the objective function table which, in turn, is calculated based on the product of the decision variables and form difficulty in columns I and J.

Step 2: Specifying and solving the test assembly model

Once the mathematical model is specified in Excel, the Opensolver parameter interface, shown in Figure 2, is used to structure and execute the ATA analysis. The user must specify the objective function by placing the cell containing the objective, Cell $Q$13[2], into the Objective Cell box. Then, the user clicks the appropriate radio button to decide whether the objective function should be maximized (max), minimized (min), or set to a specific target (value of). In our example, we click minimize because the goal is to create two forms with the same mean test score. The decision variables table ($F$3:$G$102[3] from Figure 1) are placed into the Variable Cells box. After the decision variables have been specified, the user adds constraints to the Constraints box using the Add Constraints button on the right side of the interface.

[2] A $ in Excel denotes a fixed cell. In this example, $Q$13 means that the fixed value in this cell location is used in the analysis. For problem 1, the value of $Q$13 is 0.00.

3 A colon in Excel describes a range of values. $F$3:$G$102 means all of the fixed values in the range that starts in cell $F$3 and ends in cell G$102.



OpenSolver - Model

What is AutoModel?

AutoModel is a feature of OpenSolver that tries to automatically determine the problem you are trying to optimise by observing the structure of the spreadsheet. It will turn its best guess into a Solver model, which you can then edit in this window.

Objective Cell: $Q$13 ○ maximise ● minimise ○ target value: 0

Variable Cells: $F$3:$G$102

Constraints:

<Add new constraint>
$N$3:$N$102 <= $M$3:$M$102
$R$3:$R$7 = $Q$3:$Q$7
$S$3:$S$7 = $Q$3:$Q$7
$F$3:$G$102 bin

☑ Make unconstrainted variable cells non-negative

☑ Show named ranges

Sensitivity Analysis □ List sensitivity analysis on the same sheet with top left cell:
□ Output sensitivity analysis: ● updating any previous output sheet ○ on a new sheet

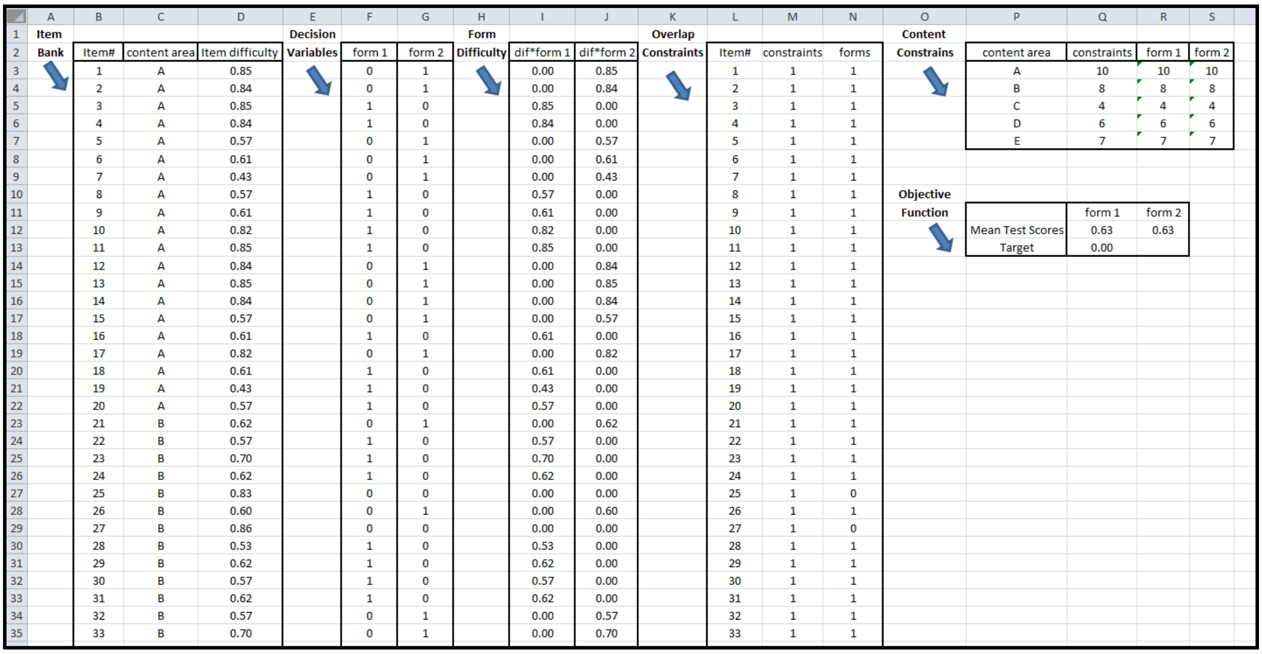
Solver Engine: Current Solver Engine: CBC

☑ Show model after saving

Figure 2. An Example of the Opensolver Parameter Interface.

Problem 1 has three sets of constraints. The first constraint guarantees that items are only used once in each parallel form (i.e., $N$3:$N$102<=$M$3:$M$102 from Figure 1). The second constraint ensures that each form satisfies the content specification (i.e., $R$3:$R$7=$Q$3:$Q$7 and $S$3:$S$7= $Q$3:$M$7 from Figure 1). The third constraint requires the decision variables to be binary (i.e., $F$3:$G$102 bin from Figure 1). The sensitivity analysis provides information about the consequence of changing the objective function for a given constraint and for adjusting a constraint that is currently zero. To ensure our example is relatively simple and straightforward, this optional analysis is not used in our demonstration. Once the model is created, the user clicks the Solver Engine button on the Solver Engine line to run the analysis and solve the optimization problem.

The result from the Opensolver analysis is an Excel spreadsheet that provides a comprehensive summary of the solution, as shown in Figure 3. Recall, Figure 1 defines the ATA problem. Figure 3 presents the solution to the problem specified in Figure 1. Hence, Figures 1 and 3 should be interpreted together. The solution contains the original item bank. The decision variable table defines the assignment of each item to each form. For problem 1, items are assigned to either form 1 or 2. A 1 means the item is assigned to a form and a 0 means the item is not assigned to the form. The form difficulty table identifies the difficulty level for each item on each form. The item overlap table defines the form assignment for the items. For example, items 1 to 24 are assigned to either form 1 or 2. But items 25 and 27 are not selected and hence not assigned to either form in this example (i.e., they receive a 0 in the overlap table). The content constraints table provides a summary of how well each form satisfies the content specification. In our example, all of the content specifications were met. The objective function table provides the mean test score for each form based on the item statistic information available in the bank. The mean test score in our example for Form 1 and 2 is 0.63 and 0.63, respectively. The means scores are computed from the item difficulty values in the bank. They indicate that the difficulty of the two forms is the same. Hence, the mean test score outcome produced by the ATA algorithm and reported in Figure 3 serves as a summary of the equivalency of the two test forms. The implication of this outcome is important for interpreting student test performance: When form 1 and 2 are randomly administered to two groups of students, the mean score on each forms is expected to be the same (barring some sampling error) because the forms were created and demonstrated to be equivalent to one another. Taken together, the results in Figure 3 identify the items that must be selected from the bank in order to create two forms that satisfy the content and statistical requirements described in our example ATA problem. The results in this figure also reveal that a feasible solution that meets the overlap, content, and statistical requirements was found for problem 1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**Item**				**Decision**			**Form**			**Overlap**				**Content**				
2	**Bank**	Item#	content area	Item difficulty	**Variables**	form 1	form 2	**Difficulty**	dif*form 1	dif*form 2	**Constraints**	Item#	constraints	forms	**Constrains**	content area	constraints	form 1	form 2
3		1	A	0.85		0	1		0.00	0.85		1	1	1		A	10	10	10
4		2	A	0.84		0	1		0.00	0.84		2	1	1		B	8	8	8
5		3	A	0.85		1	0		0.85	0.00		3	1	1		C	4	4	4
6		4	A	0.84		1	0		0.84	0.00		4	1	1		D	6	6	6
7		5	A	0.57		0	1		0.00	0.57		5	1	1		E	7	7	7
8		6	A	0.61		0	1		0.00	0.61		6	1	1					
9		7	A	0.43		0	1		0.00	0.43		7	1	1					
10		8	A	0.57		1	0		0.57	0.00		8	1	1	**Objective**				
11		9	A	0.61		1	0		0.61	0.00		9	1	1	**Function**		form 1	form 2	
12		10	A	0.82		1	0		0.82	0.00		10	1	1		Mean Test Scores	0.63	0.63	
13		11	A	0.85		1	0		0.85	0.00		11	1	1		Target	0.00		
14		12	A	0.84		0	1		0.00	0.84		12	1	1					
15		13	A	0.85		0	1		0.00	0.85		13	1	1					
16		14	A	0.84		0	1		0.00	0.84		14	1	1					
17		15	A	0.57		0	1		0.00	0.57		15	1	1					
18		16	A	0.61		1	0		0.61	0.00		16	1	1					
19		17	A	0.82		0	1		0.00	0.82		17	1	1					
20		18	A	0.61		1	0		0.61	0.00		18	1	1					
21		19	A	0.43		1	0		0.43	0.00		19	1	1					
22		20	A	0.57		1	0		0.57	0.00		20	1	1					
23		21	B	0.62		0	1		0.00	0.62		21	1	1					
24		22	B	0.57		1	0		0.57	0.00		22	1	1					
25		23	B	0.70		1	0		0.70	0.00		23	1	1					
26		24	B	0.62		1	0		0.62	0.00		24	1	1					
27		25	B	0.83		0	0		0.00	0.00		25	1	0					
28		26	B	0.60		0	1		0.00	0.60		26	1	1					
29		27	B	0.86		0	0		0.00	0.00		27	1	0					
30		28	B	0.53		1	0		0.53	0.00		28	1	1					
31		29	B	0.62		1	0		0.62	0.00		29	1	1					
32		30	B	0.57		1	0		0.57	0.00		30	1	1					
33		31	B	0.62		1	0		0.62	0.00		31	1	1					
34		32	B	0.57		0	1		0.00	0.57		32	1	1					
35		33	B	0.70		0	1		0.00	0.70		33	1	1					

Figure 3. The Solution for Problem 1.

Two Practical Problems Using Test Data from an Undergraduate Psychology Course

We began with a structured example to illustrate the logic required to build a model and solve a ATA problem using simulated data. Next, we apply this method in two practical testing situations using actual student response data. An item bank containing 144 multiple-choice items was used. The data for the two practical examples in this section of our paper were collected from a midterm and a final multiple-choice exam administered in an introductory undergraduate course focused on adolescent developmental psychology. The content codes for the items were created by the instructor of the course. The item response data were collected from all 162 undergraduate students who completed both exams during the Winter 2016 semester.

Problem 2: Parallel Forms Construction with Common Overlapping Items

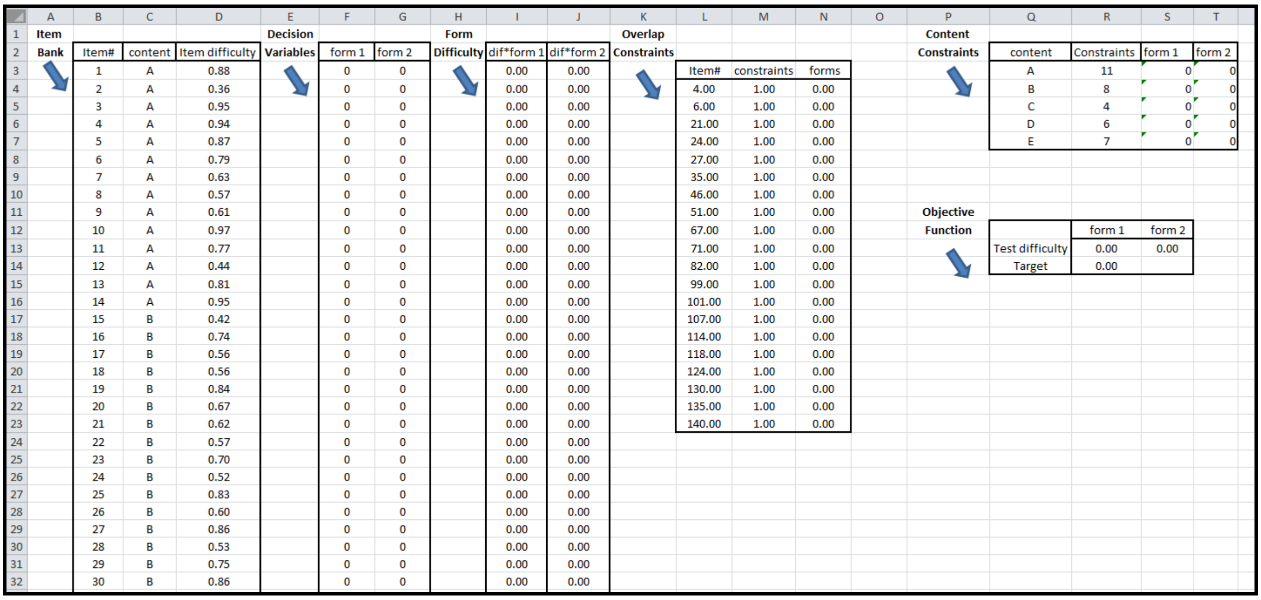
In the second problem, the goal is to create two parallel forms that meet the following requirements: (1) each form should contain 36 items; (2) the forms should have a similar mean test score; (3) the forms must meet five content areas requirements—11 items are from content A, eight items are from content B, four items are from content C, six items are from content D, seven items are from content E; and (4) the forms must contain some common, overlapping items. Item overlap is included as a constraint in problem 2 because the item bank is relatively small. To address this limitation, a unique set of 20 items that measure key concepts from each content area in the course was first identified by the instructor. These key items were only used once in the test assembly problem to limit their exposure. The remaining items in the bank were free to vary meaning that they could be used in neither form or on either one or both forms. A summary of the item bank and the ATA requirements for problem 2 are presented in Table 2.

Table 2. Summary of Content and Statistical Requirements for Test Assembly Problem 2

Content Area	Number of Items in Bank	Difficulty Mean (SD)	Number of Items Per Form
A	14	0.75(0.20)	11
B	30	0.66(0.16)	8
C	15	0.74(0.17)	4
D	53	0.72(0.17)	6
E	32	0.67(0.15)	7
Total	144*	0.71 (0.17)	36

*Items 4, 6, 21, 24, 27, 35, 46, 51, 67, 71, 82, 99, 101, 107, 114, 118, 124, 130, 135 and 140 can only be used in one of the two forms.

The mathematical model for problem 2 is presented in Figure 4 (this figure shows 32 out of 144 items). The structure of problem 2 is comparable to problem 1, except the overlap constraints are define for a specific number of items and the proportion of items across the five content areas is defined to satisfy the problem requirements.



Item Bank | Decision Variables | Form Difficulty

Item#	content	Item difficulty	form 1	form 2	dif*form 1	dif*form 2
1	A	0.88	0	0	0.00	0.00
2	A	0.36	0	0	0.00	0.00
3	A	0.95	0	0	0.00	0.00
4	A	0.94	0	0	0.00	0.00
5	A	0.87	0	0	0.00	0.00
6	A	0.79	0	0	0.00	0.00
7	A	0.63	0	0	0.00	0.00
8	A	0.57	0	0	0.00	0.00
9	A	0.61	0	0	0.00	0.00
10	A	0.97	0	0	0.00	0.00
11	A	0.77	0	0	0.00	0.00
12	A	0.44	0	0	0.00	0.00
13	A	0.81	0	0	0.00	0.00
14	A	0.95	0	0	0.00	0.00
15	B	0.42	0	0	0.00	0.00
16	B	0.74	0	0	0.00	0.00
17	B	0.56	0	0	0.00	0.00
18	B	0.56	0	0	0.00	0.00
19	B	0.84	0	0	0.00	0.00
20	B	0.67	0	0	0.00	0.00
21	B	0.62	0	0	0.00	0.00
22	B	0.57	0	0	0.00	0.00
23	B	0.70	0	0	0.00	0.00
24	B	0.52	0	0	0.00	0.00
25	B	0.83	0	0	0.00	0.00
26	B	0.60	0	0	0.00	0.00
27	B	0.86	0	0	0.00	0.00
28	B	0.53	0	0	0.00	0.00
29	B	0.75	0	0	0.00	0.00
30	B	0.86	0	0	0.00	0.00

Overlap Constraints

Item#	constraints	forms
4.00	1.00	0.00
6.00	1.00	0.00
21.00	1.00	0.00
24.00	1.00	0.00
27.00	1.00	0.00
35.00	1.00	0.00
46.00	1.00	0.00
51.00	1.00	0.00
67.00	1.00	0.00
71.00	1.00	0.00
82.00	1.00	0.00
99.00	1.00	0.00
101.00	1.00	0.00
107.00	1.00	0.00
114.00	1.00	0.00
118.00	1.00	0.00
124.00	1.00	0.00
130.00	1.00	0.00
135.00	1.00	0.00
140.00	1.00	0.00

Content Constraints

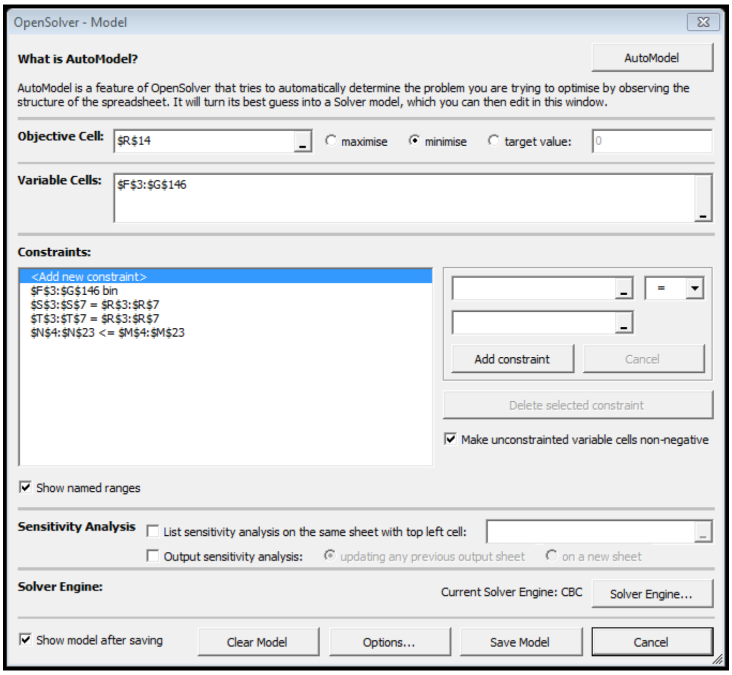
content	Constraints	form 1	form 2
A	11	0	0
B	8	0	0
C	4	0	0
D	6	0	0
E	7	0	0

Objective Function

	form 1	form 2
Test difficulty	0.00	0.00
Target	0.00	

Figure 4. Model Specified for the ATA Problem in Problem 2.

Figure 5 contains the Opensolver parameter interface. This interface contains the input required to structure and execute the ATA analysis. It includes a definition for the objective function, the decision variables, and the user-defined constraints.



OpenSolver - Model

What is AutoModel?

AutoModel is a feature of OpenSolver that tries to automatically determine the problem you are trying to optimise by observing the structure of the spreadsheet. It will turn its best guess into a Solver model, which you can then edit in this window.

Objective Cell: $R$14 ○ maximise ● minimise ○ target value: 0

Variable Cells: $F$3:$G$146

Constraints:

```
<Add new constraint>
$F$3:$G$146 bin
$S$3:$S$7 = $R$3:$R$7
$T$3:$T$7 = $R$3:$R$7
$N$4:$N$23 <= $M$4:$M$23
```

☑ Make unconstrainted variable cells non-negative

☑ Show named ranges

Sensitivity Analysis □ List sensitivity analysis on the same sheet with top left cell:

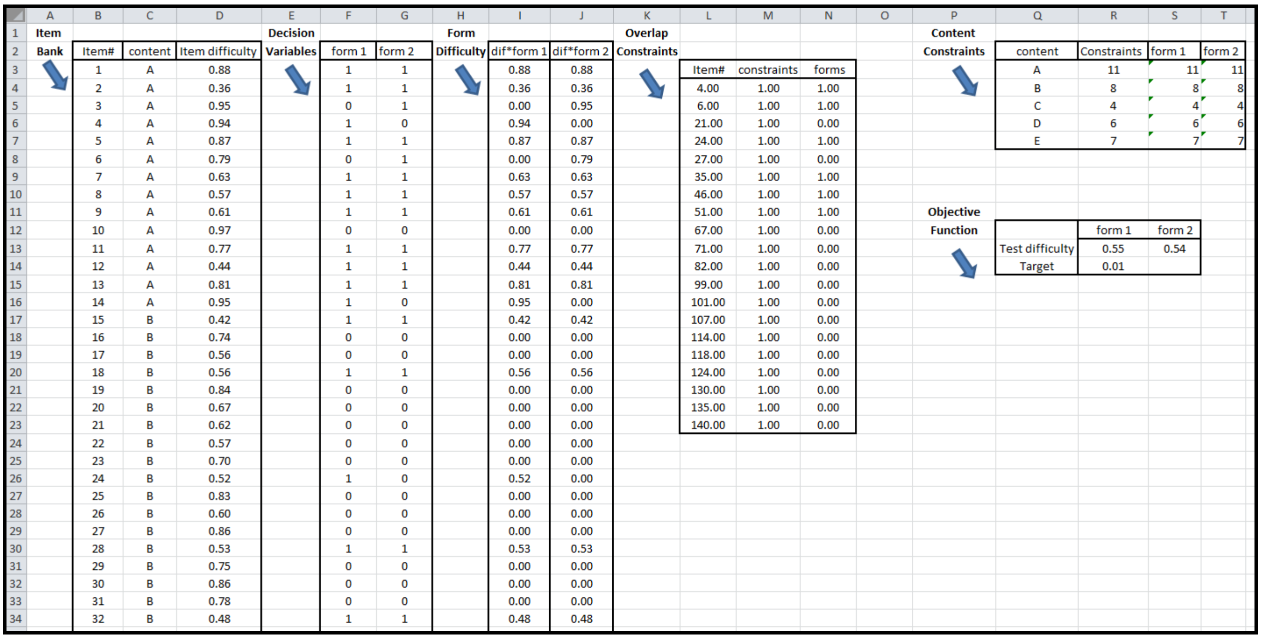
□ Output sensitivity analysis: ● updating any previous output sheet ○ on a new sheet

Solver Engine: Current Solver Engine: CBC

☑ Show model after saving

Figure 5. The Opensolver Parameter Interface for Problem 2.

The solution for problem 2 is presented in Figure 6. The solution contains the original item bank with items 1 to 144 (this figure shows 32 out of 144 items). The decision variable table defines the assignment of each item to each form. The form difficulty table identifies the difficulty level for each item. The item overlap table defines the form assignment for the items. It also identifies which items will and will not be used from the bank. The content constraints table provides a summary of how well each form satisfies the content specification. The objective function table provides the mean test score for each form. The mean test score for Form 1 and 2 is 0.55 and 0.54, respectively, using the student response data available in the bank. In other words, for problem 2, the best solution that could be produced given the required constraints is that the mean score is very similar, but not identical, across the two forms (difference between the mean scores across the two forms, as reported in Objective Function summary is 0.01). Form 2 is more difficult than form 1 by 1 score point.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Item				Decision			Form			Overlap					Content				
2	Bank	Item#	content	Item difficulty	Variables	form 1	form 2	Difficulty	dif*form 1	dif*form 2	Constraints					Constraints	content	Constraints	form 1	form 2
3		1	A	0.88		1	1		0.88	0.88		Item#	constraints	forms			A	11	11	11
4		2	A	0.36		1	1		0.36	0.36		4.00	1.00	1.00			B	8	8	8
5		3	A	0.95		0	1		0.00	0.95		6.00	1.00	1.00			C	4	4	4
6		4	A	0.94		1	0		0.94	0.00		21.00	1.00	0.00			D	6	6	6
7		5	A	0.87		1	1		0.87	0.87		24.00	1.00	1.00			E	7	7	7
8		6	A	0.79		0	1		0.00	0.79		27.00	1.00	0.00						
9		7	A	0.63		1	1		0.63	0.63		35.00	1.00	1.00						
10		8	A	0.57		1	1		0.57	0.57		46.00	1.00	1.00						
11		9	A	0.61		1	1		0.61	0.61		51.00	1.00	1.00		Objective				
12		10	A	0.97		0	0		0.00	0.00		67.00	1.00	0.00		Function		form 1	form 2	
13		11	A	0.77		1	1		0.77	0.77		71.00	1.00	0.00			Test difficulty	0.55	0.54	
14		12	A	0.44		1	1		0.44	0.44		82.00	1.00	0.00			Target	0.01		
15		13	A	0.81		1	1		0.81	0.81		99.00	1.00	0.00						
16		14	A	0.95		1	0		0.95	0.00		101.00	1.00	0.00						
17		15	B	0.42		1	1		0.42	0.42		107.00	1.00	0.00						
18		16	B	0.74		0	0		0.00	0.00		114.00	1.00	0.00						
19		17	B	0.56		0	0		0.00	0.00		118.00	1.00	0.00						
20		18	B	0.56		1	1		0.56	0.56		124.00	1.00	0.00						
21		19	B	0.84		0	0		0.00	0.00		130.00	1.00	0.00						
22		20	B	0.67		0	0		0.00	0.00		135.00	1.00	0.00						
23		21	B	0.62		0	0		0.00	0.00		140.00	1.00	0.00						
24		22	B	0.57		0	0		0.00	0.00										
25		23	B	0.70		0	0		0.00	0.00										
26		24	B	0.52		1	0		0.52	0.00										
27		25	B	0.83		0	0		0.00	0.00										
28		26	B	0.60		0	0		0.00	0.00										
29		27	B	0.86		0	0		0.00	0.00										
30		28	B	0.53		1	1		0.53	0.53										
31		29	B	0.75		0	0		0.00	0.00										
32		30	B	0.86		0	0		0.00	0.00										
33		31	B	0.78		0	0		0.00	0.00										
34		32	B	0.48		1	1		0.48	0.48										

Figure 6. The Solution for Problem 2.

Problem 3: Parallel forms construction with practice test items

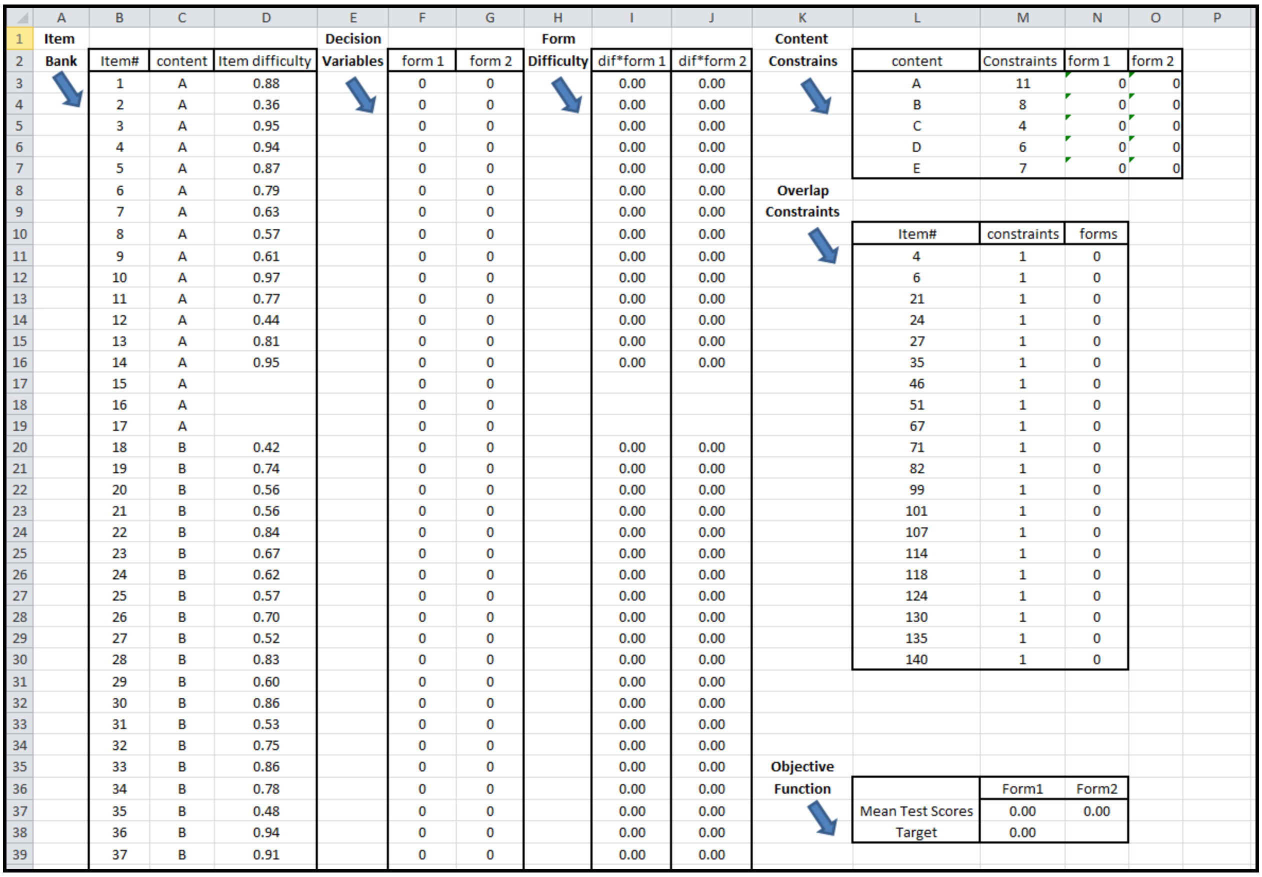
In the third problem, the goal is to create two parallel forms that meet the following requirements: (1) each form should contain 36 items; (2) the forms should have a similar mean test score; (3) the forms must meet five content areas requirements—11 items are from content A, eight items are from content B, four items are from content C, six items are from content D, seven items are from content E; (4) the forms must contain a specific set of common items (i.e., the two forms contain some overlapping test items), and (5) the forms include nine new items that do not contain item statistics. The fifth constraint is included because the instructor has written nine new items and these items have not been administered to students. Hence, the purpose is to include these new items across the two forms so the final forms satisfy the content and overlap constraints with the added benefit of collecting statistical item analysis data on the new items so they can be used in future test assembly tasks. Constraint 5 is often included in ATA problems when the long-term goal is to increase the size of the item bank. The same bank of 144 test items from problem 2 was used in problem 3, but with the addition of nine new items to produce a bank of 153 items. A summary of the item bank along with the ATA requirements are presented in Table 3.

Table 3. Summary of Content and Statistical Requirements for Test Assembly Problem 3

Content Area	Number of Items in Bank	Difficulty Mean (SD)	Number of Items Per Form
A	17(14+3 new)	0.75 (0.20)	11
B	30 (30 + 0 new)	0.66 (0.16)	8
C	17(15+2 new)	0.74 (0.17)	4
D	55(53+2 new)	0.72 (0.17)	6
E	34(32+ 2 new)	0.67 (0.15)	7
Total	153*	0.71 (0.17)	36

*****Items 4, 6, 21, 24, 27, 35, 46, 51, 67, 71, 82, 99, 101, 107, 114, 118, 124, 130, 135 and 140 can only be used in one of the two forms.

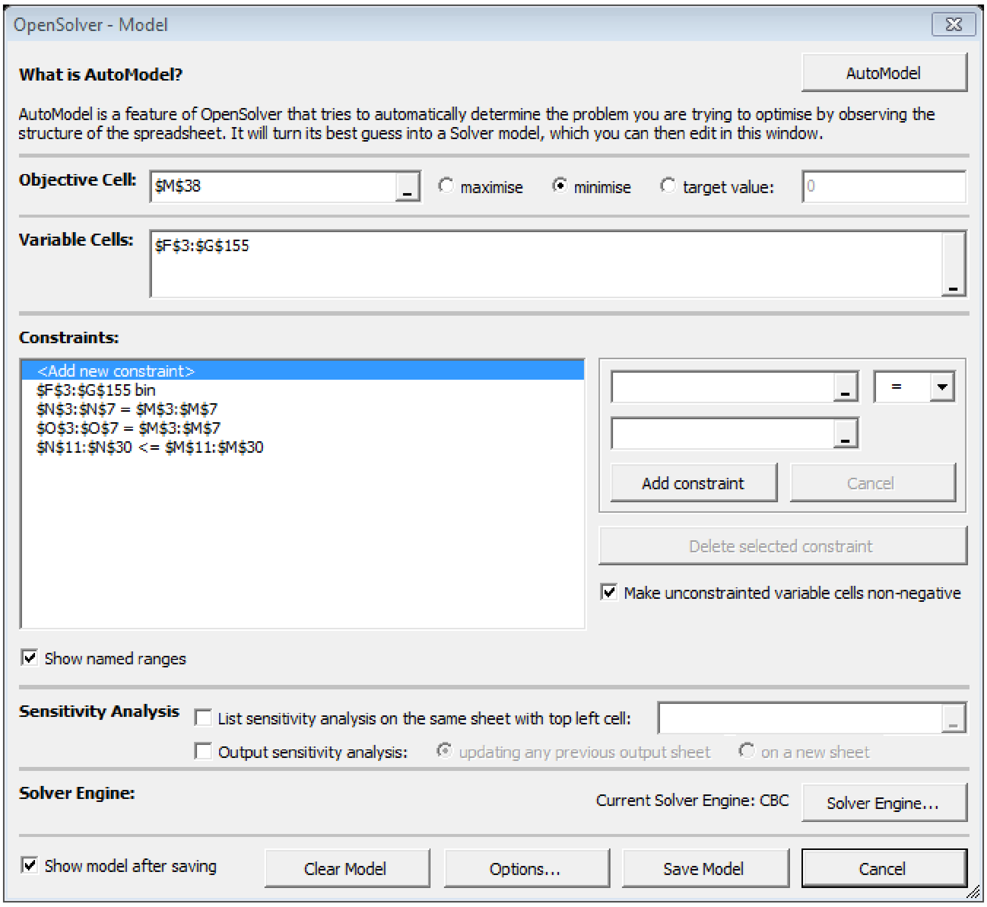
The model for problem 3 is presented in Figure 7. The structure of problem 3 is similar to problem 2, except we add nine new items which have no difficulty statistics (e.g., items 15-17 are new therefore difficulty level is blank).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Item				Decision			Form			Content					
2	Bank	Item#	content	Item difficulty	Variables	form 1	form 2	Difficulty	dif*form 1	dif*form 2	Constrains	content	Constraints	form 1	form 2	
3		1	A	0.88		0	0		0.00	0.00		A	11	0	0	
4		2	A	0.36		0	0		0.00	0.00		B	8	0	0	
5		3	A	0.95		0	0		0.00	0.00		C	4	0	0	
6		4	A	0.94		0	0		0.00	0.00		D	6	0	0	
7		5	A	0.87		0	0		0.00	0.00		E	7	0	0	
8		6	A	0.79		0	0		0.00	0.00	Overlap					
9		7	A	0.63		0	0		0.00	0.00	Constraints					
10		8	A	0.57		0	0		0.00	0.00		Item#	constraints	forms		
11		9	A	0.61		0	0		0.00	0.00		4	1	0		
12		10	A	0.97		0	0		0.00	0.00		6	1	0		
13		11	A	0.77		0	0		0.00	0.00		21	1	0		
14		12	A	0.44		0	0		0.00	0.00		24	1	0		
15		13	A	0.81		0	0		0.00	0.00		27	1	0		
16		14	A	0.95		0	0		0.00	0.00		35	1	0		
17		15	A			0	0					46	1	0		
18		16	A			0	0					51	1	0		
19		17	A			0	0					67	1	0		
20		18	B	0.42		0	0		0.00	0.00		71	1	0		
21		19	B	0.74		0	0		0.00	0.00		82	1	0		
22		20	B	0.56		0	0		0.00	0.00		99	1	0		
23		21	B	0.56		0	0		0.00	0.00		101	1	0		
24		22	B	0.84		0	0		0.00	0.00		107	1	0		
25		23	B	0.67		0	0		0.00	0.00		114	1	0		
26		24	B	0.62		0	0		0.00	0.00		118	1	0		
27		25	B	0.57		0	0		0.00	0.00		124	1	0		
28		26	B	0.70		0	0		0.00	0.00		130	1	0		
29		27	B	0.52		0	0		0.00	0.00		135	1	0		
30		28	B	0.83		0	0		0.00	0.00		140	1	0		
31		29	B	0.60		0	0		0.00	0.00						
32		30	B	0.86		0	0		0.00	0.00						
33		31	B	0.53		0	0		0.00	0.00						
34		32	B	0.75		0	0		0.00	0.00						
35		33	B	0.86		0	0		0.00	0.00	Objective					
36		34	B	0.78		0	0		0.00	0.00	Function		Form1	Form2		
37		35	B	0.48		0	0		0.00	0.00		Mean Test Scores	0.00	0.00		
38		36	B	0.94		0	0		0.00	0.00		Target	0.00			
39		37	B	0.91		0	0		0.00	0.00						

Figure 7. Model Specified for the ATA Problem in Problem 3.

Figure 8 contains the Opensolver parameter interface. This interface contains the input required to structure and execute the ATA analysis. It includes a definition for the objective function, the decision variables, and the user-defined constraints.



OpenSolver - Model

What is AutoModel?

AutoModel is a feature of OpenSolver that tries to automatically determine the problem you are trying to optimise by observing the structure of the spreadsheet. It will turn its best guess into a Solver model, which you can then edit in this window.

Objective Cell: $M$38 ○ maximise ● minimise ○ target value: 0

Variable Cells: $F$3:$G$155

Constraints:

<Add new constraint>
$F$3:$G$155 bin
$N$3:$N$7 = $M$3:$M$7
$O$3:$O$7 = $M$3:$M$7
$N$11:$N$30 <= $M$11:$M$30

☑ Make unconstrainted variable cells non-negative

☑ Show named ranges

Sensitivity Analysis □ List sensitivity analysis on the same sheet with top left cell:

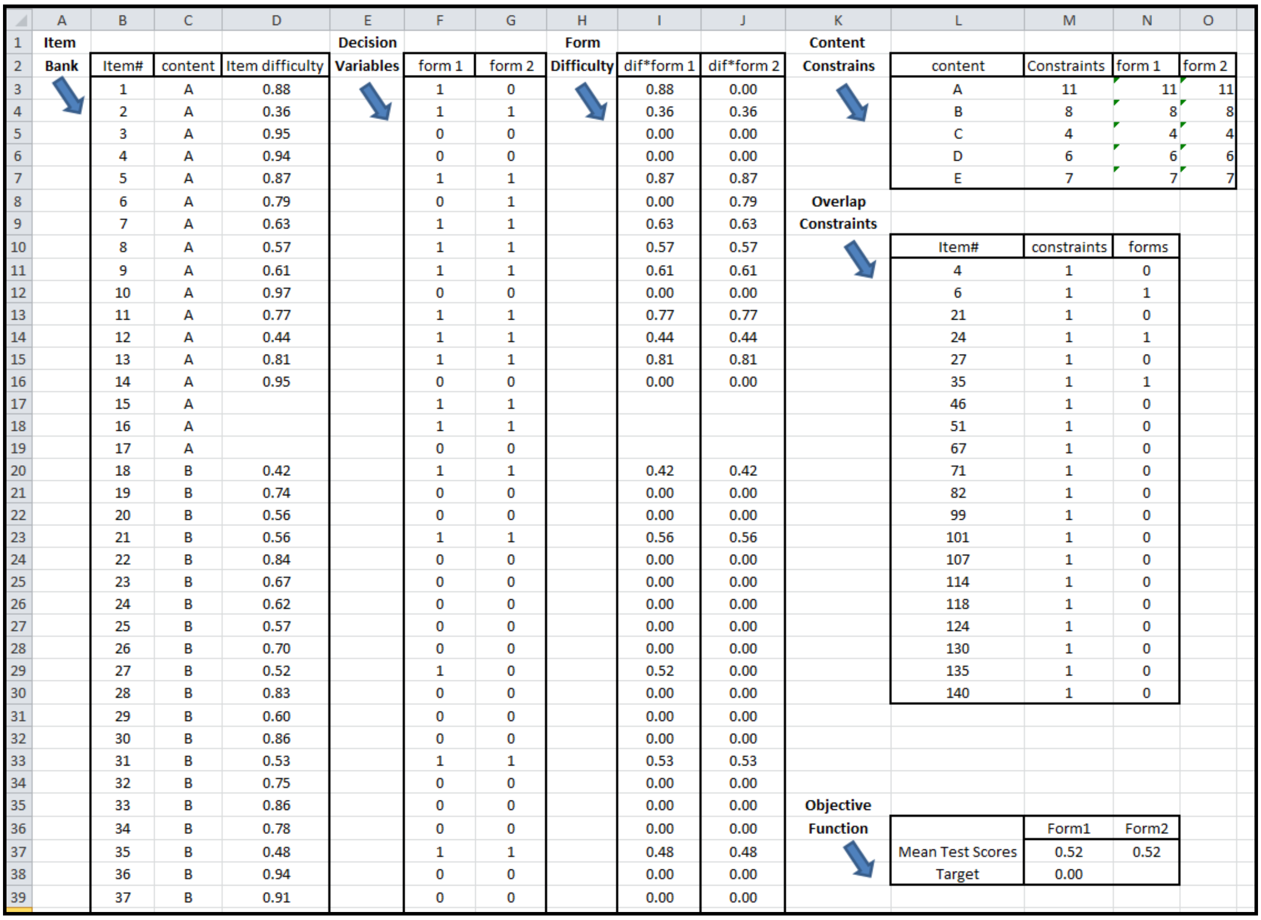
□ Output sensitivity analysis: ● updating any previous output sheet ○ on a new sheet

Solver Engine: Current Solver Engine: CBC

☑ Show model after saving

Figure 8. The Opensolver Parameter Interface for Problem 3.

The solution for problem 3 is presented in Figure 9. The solution contains the original item bank as well as the decision variable, the form difficulty, the item overlap, the content constraints, and the objective function tables. For problem 3, the mean test score for Form 1 and 2 is 0.52 and 0.52, respectively, using the student response data available in the bank. The decision variable table specifies the items to include in each form so the content constrain is satisfied. These forms will contain both the original and the new items.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Item				Decision			Form			Content				
2	Bank	Item#	content	Item difficulty	Variables	form 1	form 2	Difficulty	dif*form 1	dif*form 2	Constrains	content	Constraints	form 1	form 2
3		1	A	0.88		1	0		0.88	0.00		A	11	11	11
4		2	A	0.36		1	1		0.36	0.36		B	8	8	8
5		3	A	0.95		0	0		0.00	0.00		C	4	4	4
6		4	A	0.94		0	0		0.00	0.00		D	6	6	6
7		5	A	0.87		1	1		0.87	0.87		E	7	7	7
8		6	A	0.79		0	1		0.00	0.79	Overlap				
9		7	A	0.63		1	1		0.63	0.63	Constraints				
10		8	A	0.57		1	1		0.57	0.57		Item#	constraints	forms	
11		9	A	0.61		1	1		0.61	0.61		4	1	0	
12		10	A	0.97		0	0		0.00	0.00		6	1	1	
13		11	A	0.77		1	1		0.77	0.77		21	1	0	
14		12	A	0.44		1	1		0.44	0.44		24	1	1	
15		13	A	0.81		1	1		0.81	0.81		27	1	0	
16		14	A	0.95		0	0		0.00	0.00		35	1	1	
17		15	A			1	1					46	1	0	
18		16	A			1	1					51	1	0	
19		17	A			0	0					67	1	0	
20		18	B	0.42		1	1		0.42	0.42		71	1	0	
21		19	B	0.74		0	0		0.00	0.00		82	1	0	
22		20	B	0.56		0	0		0.00	0.00		99	1	0	
23		21	B	0.56		1	1		0.56	0.56		101	1	0	
24		22	B	0.84		0	0		0.00	0.00		107	1	0	
25		23	B	0.67		0	0		0.00	0.00		114	1	0	
26		24	B	0.62		0	0		0.00	0.00		118	1	0	
27		25	B	0.57		0	0		0.00	0.00		124	1	0	
28		26	B	0.70		0	0		0.00	0.00		130	1	0	
29		27	B	0.52		1	0		0.52	0.00		135	1	0	
30		28	B	0.83		0	0		0.00	0.00		140	1	0	
31		29	B	0.60		0	0		0.00	0.00					
32		30	B	0.86		0	0		0.00	0.00					
33		31	B	0.53		1	1		0.53	0.53					
34		32	B	0.75		0	0		0.00	0.00					
35		33	B	0.86		0	0		0.00	0.00	Objective				
36		34	B	0.78		0	0		0.00	0.00	Function		Form1	Form2	
37		35	B	0.48		1	1		0.48	0.48		Mean Test Scores	0.52	0.52	
38		36	B	0.94		0	0		0.00	0.00		Target	0.00		
39		37	B	0.91		0	0		0.00	0.00					

Figure 9. The Solution for Problem 3.

DISCUSSION and CONCLUSION

The purpose of our study was to describe and illustrate a test development process that can be used for parallel forms construction using the selected-response item type in order to permit on-demand testing. We presented the logic and highlighted the benefits of parallel forms construction using three different examples. In each example, parallel forms were created that met strict content and statistical constraints. These findings are important because educational testing is undergoing profound changes spurred on, in part, by the application of technology to assessment. As a result, CBT has become commonplace across all levels of education ranging from K-12 to post-secondary levels. This expansion is fueled by the benefits of CBT over paper-based testing that includes the ability to test on-demand, to administer exams at different locations and with different technologies, to provide students with instant feedback while relieving instructors from the monotonous task of manual scoring, and to use dynamic new item types.

Despite these important benefits, exam security remains an important concern. When computer-based tests are administered more frequently, item exposure rates will also increase. To address this concern, parallel test forms are created. Parallel forms are considered to be secure exams because each form contains a different set of items thereby minimizing item exposure in order to enhance test security. But manually assembling parallel test forms is impractical because it requires the solution to a complex combinatorial problem. As an alternative, instructors can implement ATA using existing software that is readily available and relatively easy to implement. In this study we demonstrated the use of the Opensolver, which is a free open-source add-in for Microsoft Excel that can be used to solve a broad range of ATA problems. We began with a simple problem using simulated data. Then we solved two practical problems using real data from an item bank developed using the student response data from two previously administered exams in a large introductory undergraduate psychology course. The bank contained content codes and statistical indices for each test item. Across all three problems in our study, parallel forms were constructed that met stringent content specifications while satisfying strict statistical targets thereby demonstrating the feasibility and benefits of using this approach for test development.

Instructors may hesitate before attempting to implement ATA due to concerns about the size and quality of their existing item bank. To be sure, item banking is an important requirement for solving test assembly problems. Ideally, a large bank is available with items that span many content areas and include a wide range of difficulty levels. This type of bank was used to solve problem 1. But it is also important to stress that test assembly is a flexible process. In reality, most banks are small or in-development. Problems 2 and 3, which were based on real student data from an undergraduate course in psychology, illustrate how to solve ATA problems under less than ideal item banking conditions.

For problem 2, item overlap rates were needed because the item bank was too small to produce two unique non-overlapping forms that satisfied the objective function of mean score equality. Hence, some common items were included in both forms. The inclusion of common items can be considered both a strength and a weakness in test design. Item overlap has the benefit of providing the instructor with common items across the forms so the performance of students who write each form can be compared directly on the same items. Item overlap also has the drawback of exposure meaning that some items from the test are being viewed by all students in the course. Item exposure has the potential of compromising test security if the same items are used in future exams and if students disclose these items to future test takers. Problems with item exposure can be reduced by decreasing the number of common items used in parallel forms construction.

For problem 3, newly created items that did not contain statistics were included in the construction of the parallel forms. These items were included so data on their statistical performance could be collected. Then, when this statistical information is available, instructors can use these items for future test assembly tasks. The inclusion of practice items can be considered both a strength and a weakness in test design. The benefit of using new items is that statistical information can be collected. This information, in turn, can be used to build a larger item bank. The drawback of using new items is that the difficulty level is unknown and as a result the overall mean score on the parallel forms could be different when the new items are used to compute students' test scores. To resolve this potential problem, new items are often not included in students' mean score calculation. Unfortunately, this introduces the issue of how to explain to students that not all test items actually contribute to their final exam score.

One simple way to address the item banking challenges described in problems 2 and 3 is implement ATA only when a large diverse bank is available. As an alternative, the solutions we presented suggest that there are creative ways to build and replenish banks. These solutions, however, will always require different types of trade-offs and compromises. But it is important to underscore that even with small item banks, ATA can be used to create parallel forms that meet stringent content specifications and statistical targets. In other words, instructors with relatively small banks can use ATA methods to create parallel forms when the goal is to promote flexible on-demand testing while at the same time maintaining security. Because parallel forms measure the same content areas and produce tests with the same difficulty levels, instructors can compare students' test results because the scores across the forms are highly comparable, if not equivalent, to one another.

REFERENCES

Breithaupt, K., & Hare, D. (2016). Automated test assembly. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 128-141). New York: Routledge.

Daniels, L., & Gierl, M. J. (in press). The impact of immediate test score reporting on university students' achievement emotions in the context of computer-based multiple-choice exams. *Learning and Instruction.*

Drasgow, F. (2016). *Technology and testing: Improving educational and psychological measurement*. New York: Routledge.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., 471-516). Washington, DC: American Council on Education.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22,* 224-236.

Luecht, R. M. (2016). Computer-based test delivery models, data, and operational implementation issues. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 179-205). New York: Routledge.

Sireci, S., & Zenisky, A. (2016). Computerized innovative item formats: Achievement and credentialing. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of Test Development* (2nd ed., 313-334). New York: Routledge.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York: Springer.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);303-320

Comparative Analysis of Common Statistical Models Used for Value-Added Assessment of School Performance*

Okul Performansının Katma-Değerli Değerlendirilmesinde Kullanılan Yaygın İstatistik Modellerinin Karşılaştırmalı Analizi

Sedat Şen ** Seock-Ho Kim *** Allan S. Cohen ****

Abstract
The purpose of this study was to compare three popular value-added models used in measuring school effectiveness based on their distinguishing characteristics. In this study, the simple fixed effects model (SFEM) and two hierarchical models (UHLMM and AHLMM) were analyzed using value-added measures obtained from a common data set with two years standard assessment data. Value-added measures obtained from these three models were analyzed to determine the impact of the differences of each model. Correlational analyses were also conducted to see whether there were meaningful relationships among these value-added models. SFEM and UHLMM models produced very similar rank orders of school effects while SFEM and AHLMM had only a moderate correlation. Thus there was not much difference between SFEM and two HLM models in terms of the rank orders of schools.

Keywords: School effectiveness, value-added assessment, value-added models, hierarchical linear models.

Öz
Bu çalışmanın amacı, okul etkililiğini ölçmede yaygın olarak kullanılan üç katma-değerli modeli ayırt edici özelliklerine dayanarak karşılaştırmaktır. Bu çalışmada iki yıllık bir standart test verisi kullanılarak bu veriden elde edilen katma-değerli ölçümler vasıtasıyla basit sabit etki modeli (SFEM) ve iki hiyerarşik doğrusal model (UHLMM ve AHLMM) analiz edilmiştir. Bu üç modelden elde edilen katma-değer ölçümleri, her modelin farklılıklarının etkisini belirlemek için analiz edildi. Bu katma değerli modellerin sonuçları arasında anlamlı ilişki olup olmadığını görmek için korelasyon analizine başvurulmuştur. SFEM ve UHLMM modelleri okul etkilerini benzer derecede sıralarken, SFEM ve AHLMM sonuçları orta derecede bir korelasyona sahiptir. Bu nedenle, okulların sıralamasına göre SFEM ve iki HLM modelinden elde edilen sonuçlar arasında çok fazla fark bulunmamıştır.

Anahtar Kelimeler: Okul etkililiği, katma-değerli değerlendirme, katma-değerli modeller, hiyerarşik doğrusal modeller.

INTRODUCTION

Over the past few decades, there has been growing interest in the effectiveness and accountability of schools around the world. As an example, this has been the case with the U.S., especially since the adoption of the No Child Left Behind act of 2001 which requires states to measure student academic achievement and to report on progress using Adequate Yearly Progress (AYP) measures (Amrein-Beardsley, 2008). This system is based on an approach which gives rewards to schools that make contributions to students' learning and sanctions those that do not make any improvement on student test scores. Early applications of this state-wide assessment have focused on the current status of

* This study was produced from the first author's Master's thesis.

** Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa-Turkey, e-mail: sedatsen@harran.edu.tr, ORCID ID: orcid.org/0000-0001-6962-4960

*** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia, USA, e-mail: shkim@uga.edu, ORCID ID: orcid.org/0000-0002-2353-7826

**** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia, USA, e-mail: acohen@uga.edu, ORCID ID: orcid.org/0000-0002-8776-9378

Received: 15.06.2017
DOI: 10.21031/epod.321840 *Accepted: 09.08.2017*

students. The current-status approach compares different cohorts of students at a single point in time (Doran & Izumi, 2004). It simply uses the percentage of students who passed the state test at the end of the school year.

Educators recognize that a one-time test score is not always a useful way to estimate school effects on student performance. Differences among schools may be due to student and school variables that are not measured in tests but that influence test scores. Current-status methods don't take socioeconomic factors into account, for example, when assessing schools' effectiveness. Although these methods are located at the heart of the state accountability system, there are at least two reasons why they're invalid and inappropriate to use for the purpose of school comparisons.

First, students come to school with different backgrounds. In other words, there is no random assignment of students to schools (Doran & Izumi, 2004) yet the statistical methodology underlying this approach assumes random assignment. This results in making unfair comparisons between disadvantaged and advantaged schools in terms of socioeconomic status.

Second, current-status methods are cumulative. They reflect the impact of learning obtained from all previous schools on students' performance scores (Doran & Izumi, 2004) but they do not differentiate current effects from previous effects. Thus, we cannot hold only the latest school accountable for a student's good or poor test score if the student has changed schools in the past. As Ballou, Sanders, and Wright (2004) note, holding schools accountable based on mean achievement levels makes no sense, when students enter those schools with large mean differences in achievement.

It is widely accepted that status-based accountability systems are likely to be flawed, resulting in inaccurate judgments of school quality (Doran & Izumi, 2004; Tekwe et el., 2004). As the shortcomings of this method increasingly become apparent, an alternative way of assessing school effectiveness using growth models has gained acceptance. This new method focuses on the improvement students in the school made during the year. Instead of considering how cohort groups have increased in knowledge, measuring individual student progress over time from one time point to the next is more reasonable in terms of "learning," which is meant to be "change." Growth models are designed to generate estimates from these kinds of data (Doran & Izumi, 2004).

In this regard, researchers have developed a method called value-added analysis (VAA) which enables them to use individual student achievement scores over time in order to identify effective schools. As defined by Tekwe et al. (2004) "Value-added is a term used to label methods of assessment of school/teacher from one year to the next and then use that measure as the basis for a performance assessment system" (p. 31). Pioneers of VAA claim that VAA generates fairer and more accurate estimates than those generated by state tests that measure only the achievement of a single year. The primary purpose of VAA is to determine the impact of teachers or schools on the progress of their students (Raudenbush, 2004). To do this, VAA computes gain scores by taking the differences between students' scores on state tests from one year to the next (Sanders et al., 2002).

The VAA approach evaluates schools based simply on how they increased the level of their students' knowledge. The two basic ideas underlying value-added measurement are that it is calculated for each individual nested within the schools and that it is based on changes in student performance from one year to the next (Ladd & Walsh, 2002). Another advantage cited of VAA is that, unlike the current-status method, it can control the effect of confounding variables such as student and school socioeconomic status that may influence the test scores. In this way, it is an attempt to minimize the influence of experiences, privilege, and ethnicity on student performance.

In general, value-added models (VAMs) are a class of statistical model procedures that analyze students' standardized test scores over time to identify the degree to which a student's progress is a function of their own characteristics or of the characteristics of their school (Doran & Izumi, 2004). VAMs have recently received a great deal of interest from both policy makers and researchers due to a belief that these models can adequately determine how individuals are growing over time while appropriately attributing that portion of their gain scores to their schools (Sanders, & Horn, 1994; Sanders, & Horn, 1998; Sanders, Saxton, & Horn, 1997). It is an area of research in education that has achieved a significant role in shaping the school accountability system.

Several VAM approaches have been suggested by researchers. Current-status methods all rely on regression models and assume that school effects are fixed (Tekwe et al., 2004). They are also confounded with nonschool factors (Sanders, 2000), whereas VAMs require the use of more complex statistical models such as mixed models and hierarchical models which assume school effects to be random. Hanushek (1972) is generally credited as the first to use VAM methods in the accountability system. Sanders, who developed the Tennessee Value Added Assessment System (TVAAS), was the first to implement VAMs in a statewide testing system (Stewart, 2006).

According to a report by the RAND corporation (McCaffrey, Lockwood, Koretz, & Hamilton, 2003) early VAM applications (e.g., Hanushek, 1972; Murnane, 1975) primarily used fixed effects models. More recent applications, including the TVAAS layered model, have used random effects models exclusively.

Another important model is one developed by Raudenbush and Bryk (1986) and Aitkin and Longford (1986). This model relies on hierarchical linear models to measure student growth. Although there are several VAMs which are based on different statistical assumptions (Braun, 2004; Tekwe et al., 2004), the most popular has been the TVAAS (Olson, 2004). For any of these models to be useful in VAA analysis, however, the test scores must be vertically scaled (Ballou et al., 2004; Doran & Cohen, 2005). That is, the test scores must all be expressed on a common scale that extends over the time periods included in the analysis. In brief, longitudinal data, annual assessment, and vertically equated tests are said to be basic elements of VAMs. Typically, standardized assessment scores are used in VAM studies. Though no VAM has yet been obvious to be clearly superior over another, VAMs are considered to be fairer and more accurate than conventional methods (Doran & Izumi, 2004).

To date, several alternative models, ranging from simple gain scores to complex mixed models, have been suggested by researchers with regard to assessment of school effectiveness. However, there have been a limited number of studies which make comparisons among these different models (Ballou et al., 2004; McCaffrey et al., 2003; Tekwe et al., 2004). Selection of the most useful model for an accountability analysis requires determining which model is most accurate. Fortunately, a few important studies have been conducted to determine the most desirable model for computing school effects. The Journal of Educational and Behavioral Statistics published one volume solely concerning the VAA and popular VAMs (Wainer, 2004). The papers in that volume concluded that there are numerous acceptable models as opposed to only a single acceptable model.

Tekwe et al. (2004), Ballou et al. (2004), and McCaffrey et al. (2003) describe differences among VAMs. As these studies have noted, compared to other methods, VAMs are less biased and produce more precise estimates. Although there is a lack of comparative studies showing which VAM is better than the others, the LMEM model has been used frequently for accountability purposes. Ballou et al. (2004) conducted a simulation study to evaluate the TVAAS model which is based on the LMEM. Results indicated that the TVAAS uses a highly parsimonious model that omits controls for contextual factors such as SES and demographics that influence achievement.

Unlike the LMEM model, HLM models include school and student variables and attempt to control such factors by statistical adjustment (Bryk & Raudenbush, 1992). Sanders et al. (2002) noted that inclusion of these factors in HLM affects the school estimates resulting in biased measures of schools towards zero. Sanders' LMEM model does not account for these variables. That model attempts to eliminate controls for these variables by use of multiple measures on each student (Ballou et al., 2004). Sanders found that the inclusion of these factors to the model did not result in a significant difference between the two models (Ballou et al., 2004). Results of a simulation study comparing the general model, which is similar to the AHLMM, with those of a layered model which is similar to the LMEM, however, suggested that the AHLMM fit the data better than the layered model (McCaffrey et al., 2003).

Tekwe et al. (2004) found little or no benefit from use of more complex models. The simpler SFEM model provided results that were more accurate compared to estimates from the other models. Results also indicated that the AHLM model would be preferred when there is a need for controlling the effects of student and school variables estimates and that selection of one of the two models should be based on non-empirical considerations.

Although VAMs have been shown as an important tool for accountability system, a number of researchers criticized the VAMs application for determining school or teacher effectiveness. An important criticism of VAMs is that they do not yet solve the problem of randomization completely (Wiley, 2006). Another criticism of VAMs is about the precision of the value-added estimates obtained from longitutional data sets. Schochet and Chiang (2010) examined the likely system error rates for measuring teacher and school performance in the upper elementary grades using ordinary least squares (OLS) and Empirical Bayes (EB) methods applied to student test score gain data.

Similarly, Guarino, Reckase, and Wooldridge (2015) investigated the accuracy of the value-added estimates of teachers obtained from commonly used value-added models. They found that no one method accurately captures true teacher effects and classifies teachers in realistic conditions. In addition, VAM approach has been shown to be invalid when there is endogeneity which may be due to correlation between the random effect in the hierarchical model and some of its covariates (Manzi, San Martín, & Van Bellegem, 2014). Another criticism of VAMs is about the data requirements of these models. As mentioned above vertically equated test results from multiple years are basic elements of VAMs. This makes VAM useful for a single developmental scale. However, most of the VAMs cannot be used for multiple test instruments (on different scales) administered within a school year. A few researchers have discussed how to use VAMs to analyze longitudinal student achievement data obtained from multiple instruments (Green, 2010; Rivkin, Hanushek, & Kain, 2005).

There have been numerous studies that show the strengths of the VAMs over the conventional methods. However, the concern remains that simpler models are as efficient as more complex models (Doran & Fleischman, 2005). Several models introduced in VAA calculate the value-added measures based on different assumptions. SFEM and UHLMM do not account for school/non-school variables, while AHLMM attempts to control these factors by statistical adjustments. In this study, the impact of school and non-school factors are compared on school-level value-added scores using an empirical data with an eye to better understanding problems associated with model complexity. Three popular VAMs (i.e., SFEM, UHLMM, and AHLMM) were examined in this study. The models selected for the present study show similarities to a previous study conducted by Tekwe et al. (2004). Tekwe et al. (2004) have also examined the LMEM in their study in addition to the models compared in this study. LMEM was excluded from our study due to data requirements of this model.

METHOD

Instrumentation

Data for this study were taken from 2002 and 2003 statewide mathematics and reading test results of the Florida Comprehensive Assessment Test (FCAT) for Grades 6 to 8. Separate analyses were done for each grade. The FCAT is a criterion-referenced test that aims to assess student achievement in high-order cognitive skills represented in the Sunshine State Standards (Florida Department of Education, 2003) in reading, mathematics, writing, and science. The FCAT includes three types of questions: multiple choice items, graded response items, and open-ended items. FCAT scaled scores used in this study were vertically scaled, thus making them appropriate for VAA.

Sample

Separated analyses were performed for each of the grade cohorts for Grades 6, 7 and 8 in a large Florida school district with 44 secondary schools for 2002 and 2003. Only standard curriculum students were used in the analyses. Special education students with any exceptionality and students in the limited English proficiency (LEP) program for two or fewer years were excluded due to following reasons. Generally, it is impossible to collect two years of score from students with severe cognitive disabilities that are required for most of the VAMs. In addition, students with limited English cannot show real performance on state test and this may have a negative effect on the value-added measures of schools. Students whose reported ages were outside the acceptable age range for a given grade were excluded from the analyses. Listwise deletion was applied to exclude these students' information.

A total of 60,718 students were available for analyses after the exclusions: 19,611 for Grade 6, 20,433 for Grade 7, and 20,674 for Grade 8. Non-school variables for socioeconomic status and minority status were included in the data set. Socioeconomic status information was provided in the form of student's eligibility for the free-or-reduced lunch program. Minority status is a school-level variable is based on the proportion of African-American or non-African-American students in the school. Descriptive statistics based on grade and subject combination are presented in Table 1.

Table 1. Sample Size, Mean FCAT and Standard Deviation by Subject, Grade and Year, and Percent Minority and Percent Poverty in 2003 by Grade

		Reading			Math			Demographics in 2003	
		2002 score	2003 score	Change score	2002 score	2003 score	Change score	Poverty	Minority
6th	*N*	19,611	19,611	19,611	19,611	19,611	19,611	19,611	19,611
	M	1421.32	1527.89	106.57	1566.02	1581.17	15.15	73.7%	28.6%
	SD	368.52	371.85	235.62	294.80	297.80	189.48		
7th	*N*	20,433	20,433	20,433	20,433	20,433	20,433	20,433	20,433
	M	1493.98	1623.32	129.33	1554.14	1692.70	138.56	72.2%	28.4%
	SD	385.43	348.92	244.52	293.74	255.18	191.43		
8th	*N*	20,674	20,674	20,674	20,674	20,674	20,674	20,674	20,674
	M	1606.93	1782.10	175.16	1675.76	1804.40	128.64	70.3%	28.6%
	SD	345.79	276.42	223.87	274.60	216.95	169.142		

Value-Added Models Used in This Study

As noted above, VAMS have the capability of controlling the effects of non-school variables as well as prior performance. In this study, results for three commonly used VAMs were compared: a simple fixed effects model and two hierarchical linear models. It should be noted that layered mixed effects model (LMEM) is another popular VAM that is useful for data sets collected from students attending multiple schools. This model was not examined in this study as the data set in this study does not have students attending multiple schools within a school-year. This makes present study different from Tekwe et al. (2004).

Simple fixed effects model (SFEM)

Fixed effects models (FEM) used for VAA assume school effects to be fixed rather than random. These have the advantage of being the simplest VAM, requiring less computation than the others. As a result, estimates from FEM are more easily understood by policymakers and educators with little statistics experience (Wiley, 2006). The simple fixed effects model (SFEM) is an extension of the FEM. One concern with this model is that it does not incorporate student-level covariates and does not apportion variance for students who have attended multiple schools. Thus it does not produce any shrunken estimates. As SFEM uses only two years of data in a single subject, however, its application is very straightforward.

Model parameterization:

$$d_{ijs} = \beta_{0s} + \sum_{k=1}^{44} \beta_{1ks} S_{kij2} + \varepsilon_{ijs} , \qquad (1)$$

where

$d_{ijs} = \gamma_{ijs2} - \gamma_{ijs1}$,

$d_{ijs} =$ is a simple change score obtained from difference between two examinations of a student i in school j on the same subject area s,

γ_{ijst} =is the test score on the subject area s $(s = 1, 2)$ at time t $(t = 1, 2)$ for the student j $(j = 1, \cdots, n_j)$ in school i $(i = 1, \cdots, n_i)$,

S_{kij2} = is effect coding at time (t = 2) for school k $(k = 1, \cdots, 44)$ with coding numbers m $(m = 1, \cdots, 43)$,

S_{kij2} = 1 for $k = m$ and $k \neq 44$; 0 for $k \neq m$ and $k \neq 44$; -1 for $k = 44$,

and ε_{ijs} is the random error for student j in school i for subject area s.

It is assumed that $\varepsilon_{ijs} \sim N(0, \sigma^2_{\varepsilon s})$.

β_{1ks} in Equation 1 is the value-added component in subject area s for school k.

Hierarchical linear models.

Hierarchical linear models (HLM) require using hierarchically ordered nested data. The hierarchical nature of the structure is that students are considered nested within classes and classes as nested within schools. Due to the nature of the data used in education, HLM has been used extensively for analysis of school effects (Raudenbush & Bryk, 2002). HLM is a special type of the general mixed models family and can be used to obtain value-added measures. These models demand more computation than SFEM, but unlike SFEM, HLM-based models produce shrunken effects.

The HLM analysis consists of four parts as follows (Raudenbush & Bryk, 1988-1989):

i. Apportioning variation between and within units of analysis
ii. Assessing the homogeneity of regression assumption
iii. Testing for compositional effects
iv. Assessing the effect of the method

Traditional regression methods assume that individuals are independent of each other although students in the same school might have similar results when compared to students from different schools. HLM can handle this violation of the independence assumption unlike linear models.

In this study, two different types of HLM were examined, unadjusted HLM (UHLMM) with random intercept and adjusted HLM (AHLMM). The AHLMM consists of two equations called student-level and school-level models. The two-level HLM provides an analytical framework for examining the effects of schools on student outcomes. An extension of two-level model (i.e., three-level HLM) can be used to obtain value-added estimates of schools and teachers using a data set structure which has students nested within teachers and teachers nested within schools.

Unadjusted hierarchical linear model (UHLMM)

UHLMM uses unadjusted change score with random intercept. This model consists of two level HLM described by the following equations;

Student-level model:

$$d_{ijs} = \beta_{0is} + \varepsilon_{ijs},$$

where d_{ijs} is the change score defined as in Equation 1, $\beta_{0is}$is a random intercept associated with the school *i*, and ε_{ijs} is a random error.

School-level model:

$$\beta_{0is} = \gamma_{0s} + \xi_{is},$$

where γ_{0s} is the mean of the random intercepts, β_{0is}, and ξ_{is} are the random effect and random error of school *i* on the random intercept for subject area *s*. $\beta_{0is}$and $\xi_{is}$are assumed to be independent**.** ε_{ijs} and ξ_{is} are assumed to have normal distribution.

Single equation form:

$$d_{ijs} = \beta_{0s} + \xi_{is} + \varepsilon_{ijs}. \qquad (2)$$

Adjusted hierarchical linear model (AHLMM)

The AHLMM model is adjusted for student-level and school-level covariates.

Student-level model:

$$d_{ijs} = \beta_{0is} + \beta_{1s}\,\gamma_{ijs1} + \beta_{2s}Min_{ij} + \beta_{3s}Pov_{ij} + \varepsilon_{ijs},$$

where $d_{ijs} = \gamma_{ijs2} - \gamma_{ijs}$, β_{0is} is a random intercept associated with the school *i* and subject area *s*, Min_{ij} = an indicator of minority status (Yes or No) for student *j* in school *i*, Pov_{ij} = an indicator of poverty in which the status of a student eligible for a free-and-reduced lunch is considered to be poverty (Yes or No) for student *j* in school *i*, β_{1s}, β_{2s}, and β_{3s}, are the fixed effects of previous year's test score, minority status, and poverty on learning gain in subject area *s*, and ε_{ijs} is a random error.

School-level model:

$$\beta_{0is} = \gamma_{0s} + \gamma_{1s}Z_{1i} + \gamma_{2s}Z_{2i} + \xi_{is},$$

where Z_{1i} is the mean input score for the school *i*, Z_{2i} is the percentage of students in poverty in the school *i*, ξ_{is} is is the random error associated with the value of the random intercept for the subject area test (*s*) and the school *i* in the student level model, and the γ's are fixed effects coefficient parameters. The within and between school error terms, ε_{ijs} and ξ_{is}, are assumed to be independent.

Single equation form:

$$d_{ijs} = \gamma_{0s} + \gamma_{1s}Z_{1i} + \gamma_{2s}Z_{2i} + \beta_{1s}\gamma_{ijs1} + \beta_{2s}Min_{ij} + \beta_{3s}Pov_{ij} + \xi_{is} + \varepsilon_{ijs}, \tag{3}$$

RESULTS

Assumptions and characteristics of each of the VAMs used in this study are shown in Table 2. Thus, differences in characteristics of the models can be seen in Table 2. Interpretations of results for each model are based on distinguishing characteristics of the model. Correlations between VAM measures of schools generated from each model are given in Table 3. Schools were ranked based on their VAM estimates from different models. Correlational results provide information about the rank order of school effects generated from each model. Tables with these rankings are also presented in Appendices.

Table 2. Summary of Distinguishing Characteristics of Models

Model identifier	Dependent variable	School effects	Student-level variable	School-level variables
SFEM	Change score	Fixed	No	No
UHLMM	Change score	Random	No	No
AHLMM	Change score	Random	Yes	Yes

Note. Adapted from Tekwe et al. (2004, p.23). SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

Table 3. Table of Correlations Between Value-added Measures of the Models

	6th grade		7th grade		8th grade	
	Math	Reading	Math	Reading	Math	Reading
SFEM vs. UHLMM	.99	.99	.99	.99	.99	.99
SFEM vs. AHLMM	.75	.85	.80	.55	.73	.74
UHLMM vs. AHLMM	.75	.85	.80	.54	.73	.74

Note. SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

With respect to the assumption of school effects as random, the SFEM is the only one that accounts for school effects as fixed effects. Therefore, it is appropriate to compare the SFEM to the UHLMM. The UHLMM differs only in that it considers the school effect to be random. The most important finding that is evident in Table 3 is the very high correlation between SFEM and UHLMM value-added estimates (r = .99) in all cohorts. This suggests that the two models provide the same rank ordering of schools. Thus, it is possible to conclude that there was no difference between taking school effects as random or fixed in terms of rank order of school effects.

A second concern in measuring school effectiveness is to include school and non-school covariates in the models. Among the models in this study, only the AHLMM can take both student-level and school-level effects into account. Apart from this characteristic, the AHLMM and UHLMM are identical. As a result, we can make inferences based on the comparison of these two models. As can be seen in Table 3, there were moderate correlations ranging from .54 to .85 between AHLMM and UHLMM for the different cohorts. This indicates that the effects of including school and non-school variables in the AHLMM had a clear impact on the VAA estimates.

Another comparison with the AHLMM can be made with SFEM. This comparison will help to see the effects of employing shrinkage or including school and non-school variables in the AHLMM model. Correlations between these two models showed moderate values ranging from .55 to .85. These results suggest there is a noticeable difference between SFEM and AHLMM. Although the AHLMM is appropriate when seeking to adjust for confounding variables, the only thing we can really conclude is that there was a difference between the rank orders of schools based on these two models.

Strong correlations were observed between results generated by the SFEM and UHLMM, but much more modest correlations were observed between the AHLMM and all other models. We conclude on the basis of these results that there was not much difference between the SFEM and hierarchical models in terms of the rank order of school estimates.

Once a model is chosen, value-added measures for students can be converted to standardized grades to determine the relative performance of the teachers within each school (or attributed to each school). To obtain standardized grades, standardized value-added measures were divided by their standard errors and assigned grade point average (GPA) values using the following criteria from Tekwe et al. (2004):

If $z > 2$, then assign a grade of A and 4 growth points;
If $1 < z \leq 2$, then assign a grade of B and 3 growth points;
If $-1 < z \leq 1$, then assign a grade of C and 2 growth points;
If $-2 < z \leq -1$, then assign a grade of D and 1 growth points;
If $z \leq -2$, then assign a grade of F and 0 growth points.

Results of the standardized grade conversions are presented in Table 4.

Since grades from the SFEM and UHLMM models were found to be similar, we present only results for the SFEM and AHLMM in Table 4. Results in Table 4 suggest that large schools with higher value-added estimates tended to have lower GPA values than smaller schools with lower value-added estimates, although it was also possible that large schools with lower value-added estimates could have higher GPA values.

Individual school estimates and their rankings were obtained for each grade from three different VAMs. Only estimates for Grade 6 are presented (see Tables 5 and 6 in Appendices A and B). (Estimates for Grades 7 and 8 are available on request from the first author.). For the SFEM, estimates can be interpreted as the difference between the school specific sample average change and the average changes overall. Estimates from the UHLMM are shrunken estimates of school effects from the SFEM. These can be calculated as estimates of the best linear unbiased predictors of the random effects for each school and each grade. Value-added estimates of the AHLMM were also calculated as estimates of best linear unbiased predictors.

The ranks of the school estimates from the SFEM were similar to those of the school estimates from the UHLMM. It is interesting to note that estimates from both models were very similar. This result also suggests that there was little difference in estimating school effects as either random or fixed. Results from the AHLMM had moderate agreement with results from SFEM. Results from each of the models suggested that VAM rankings of schools differed across different grades. Results compared for each grade, however, were very consistent with the results of correlational analyses.

Table 4. Growth Point Averages for Each School Based on Value-Added Measures from SFEM and AHLMM

	SFEM						AHLMM					
School	M	R	G6	G7	G8	T	M	R	G6	G7	G8	T
1	0.00	0.66	0.50	0.50	0.00	0.33	0.33	1.33	0.50	1.50	0.50	0.83
2	3.00	3.66	2.00	4.00	4.00	3.33	2.33	2.33	1.50	2.50	3.00	2.33
3	1.00	2.33	0.50	1.00	3.50	1.66	2.33	2.66	2.00	2.00	3.50	2.5
4	2.00	1.00	2.00	0.50	2.00	1.50	2.00	1.66	2.50	1.00	2.00	1.83
5	3.33	3.00	2.00	4.00	3.50	3.16	2.66	1.66	2.00	2.50	2.00	2.16
6	2.66	1.66	2.50	2.50	1.50	2.16	2.66	2.00	3.00	2.00	2.00	2.33
7	1.00	2.33	1.50	1.00	2.50	1.66	1.66	2.00	2.00	1.50	2.00	1.83
8	3.00	2.66	4.00	2.00	2.50	2.83	1.66	2.00	2.00	1.50	2.00	1.83
9	2.00	1.00	0.00	1.00	3.50	1.50	2.33	2.00	2.00	2.00	2.50	2.16
10	3.66	3.00	4.00	3.00	3.00	3.33	3.00	2.33	2.50	2.50	3.00	2.66
11	2.33	1.66	3.50	2.50	0.00	2.00	2.33	1.33	2.00	2.50	1.00	1.83
12	3.00	2.33	3.00	2.50	2.50	2.66	1.66	1.66	2.00	1.50	1.50	1.66
13	2.66	2.00	4.00	2.00	1.00	2.33	2.33	2.33	3.00	2.00	2.00	2.33
14	1.66	2.00	4.00	1.50	0.00	1.83	2.00	2.00	2.00	2.00	2.00	2.00
15	1.00	2.33	2.00	1.00	2.00	1.66	1.00	2.00	2.00	1.00	1.50	1.50
16	0.00	1.00	1.00	0.00	0.50	0.50	1.33	2.00	1.00	2.00	2.00	1.66
17	0.33	1.00	1.50	0.00	0.50	0.66	0.66	1.33	2.00	0.50	0.50	1.00
18	0.00	1.33	0.00	0.00	2.00	0.66	2.00	2.66	2.00	2.00	3.00	2.33
19	3.33	3.00	3.50	4.00	2.00	3.16	2.66	1.66	3.00	2.50	1.00	2.16
20	1.66	1.66	1.00	2.00	2.00	1.66	2.00	2.00	2.00	2.00	2.00	2.00
21	1.00	1.00	0.00	1.50	1.50	1.00	0.66	1.00	1.00	1.50	0.00	0.83
22	2.00	2.33	2.00	0.50	4.00	2.16	2.00	2.66	2.50	1.50	3.00	2.33
23	3.00	2.00	1.50	3.00	3.00	2.50	3.66	2.33	3.00	3.00	3.00	3.00
24	1.33	2.66	1.50	3.00	1.50	2.00	0.66	2.00	2.00	1.00	1.00	1.33
25	3.33	2.66	4.00	2.50	2.50	3.00	3.33	2.33	4.00	2.00	2.50	2.83
26	1.00	1.66	2.50	1.50	0.00	1.33	0.66	1.33	1.50	1.50	0.00	1.00
27	2.00	2.33	1.50	3.50	1.00	2.16	2.33	2.66	2.50	3.50	1.50	2.50
28	1.66	2.66	3.50	1.00	2.00	2.16	1.66	2.00	2.00	1.50	2.00	1.83
29	3.66	2.66	3.00	3.00	3.50	3.16	2.00	2.00	2.00	2.00	2.00	2.00
30	0.66	2.33	0.50	1.00	3.00	1.50	2.66	3.00	2.00	2.50	4.00	2.83
31	1.66	2.66	2.00	1.50	3.00	2.16	2.66	2.66	2.50	2.00	3.50	2.66
32	2.33	3.00	1.50	2.50	4.00	2.66	3.33	3.33	3.00	3.00	4.00	3.33
33	2.00	2.66	0.00	4.00	3.00	2.33	1.33	1.66	0.00	2.50	2.00	1.50
34	1.33	1.33	4.00	0.00	0.00	1.33	2.00	2.00	2.00	2.00	2.00	2.00
35	0.00	0.66	0.00	0.00	1.00	0.33	1.33	2.00	1.00	1.50	2.50	1.66
36	2.33	1.66	1.50	2.50	2.00	2.00	2.00	1.33	1.00	2.00	2.00	1.66
37	3.33	1.33	3.00	2.50	1.50	2.33	2.66	1.33	2.50	2.50	1.00	2.00
38	2.66	0.66	1.00	2.00	2.00	1.66	2.66	1.66	1.50	2.50	2.50	2.16
39	2.66	3.33	4.00	2.50	2.50	3.00	2.00	2.33	2.00	2.00	2.50	2.16
40	2.66	2.66	1.50	3.50	3.00	2.66	1.33	1.33	1.50	1.50	1.00	1.33
41	1.33	0.66	3.00	0.00	0.00	1.00	2.66	1.66	2.50	1.50	2.50	2.16
42	2.00	1.66	2.50	2.50	0.50	1.83	2.00	2.00	1.50	2.50	2.00	2.00
43	2.00	0.66	0.00	3.00	1.00	1.33	1.00	1.00	0.50	1.50	1.00	1.00
44	-	-	-	-	-	-	2.33	2.33	2.00	2.50	2.50	2.33

Notes. M = Math GPA; R = Reading GPA; T = Total GPA; 6G = 6th Grade GPA; 7G = 7th Grade GPA.

DISCUSSION and CONCLUSION

The purpose of the present study was to determine whether there were similarities or differences among three models commonly used for value-added assessment of schools. The simplest model was the SFEM. This model treats school effects as fixed. Two hierarchical linear models were also included. Each model has distinguishing characteristics and different assumptions. Value-added estimates of individual schools obtained from these models were analyzed to compare results from the different models on the estimates.

The primary question was to investigate whether results from simpler models, such as the SFEM, differed as effective as the more complex models such as AHLMM in terms of school rankings. Previous research has found that little difference between the results of simple and complex value-added models in that correlations between estimates from SFEM and AHLMM models ranged from .55 to .85 (Tekwe et al., 2004). Results from this study were somewhat consistent with previous research in that the simple model produced similar rank orders of school effects with the more complex AHLMM. Based on these results, it may be concluded that simple models were as effective as more complex models at estimating value added effects of schooling. Further, simpler models generally could be used in place of more complex models such as AHLMM. There is typically a desire for using simpler statistical models among policy makers as well as the general public. Results of the present study tend to support the use of simpler models such as the SFEM in value-added accountability systems.

Another concern in value-added studies is to determine the impact of the inclusion of school and student background variables into models on model estimates. Among the models in this study, only the AHLMM includes statistical adjustments for these potentially confounding variables. Tekwe et al. (2004) suggested that both inclusion and exclusion of these variables during the analysis result in biased estimates of schools. In this study, the estimates from the AHLMM model were compared to estimates from other models to determine the effects of these covariates. No major differences were observed between results of the AHLMM, the UHLMM and the SFEM. Correlations between estimates from the AHLMM and SFEM ranged from .55 to .85. Correlations between results from the AHLMM and the UHLMM also ranged from .54 to .85. These correlations were mostly consistent with results from previous research. Consistent with previous research, inclusion of these covariates did have an effect on value-added estimates. The omission of covariates from the model appeared to bias parameter estimates when students were stratified by those covariates (McCaffrey et al., 2003).

The present study also reported on standardized GPA grading and rankings of each school based on value-added estimates from each model. These results were consistent with the correlational analysis. VAM-based rankings of schools showed differences over grades. It should be noted that the conclusions drawn from this study cannot be generalized to teachers or to other test conditions.

Although, value-added models are believed to be useful in school accountability system, the credibility of these methods have been questioned by a number of researchers (AERA, 2015, Amrein-Beardsley, 2014; Ballou & Springer, 2015; Guzman, 2016; The American Statistical Association (ASA), 2014). Amrein-Beardsley (2014), emphasized that VAMs have several problems with reliability, validity, and bias, affecting their fairness and transparency. In addition to these serious problems, theoretical and methodological assumptions of VAMs have also been questioned in the literature. Thus, school (or teacher) performances should not be based on only value-added measures obtained from any of the VAMs described in this study. As Amrein-Beardsley (2014) suggested multiple measures and more holistic evaluation systems should be used for school evaluations rather than relying only on VAMs.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modeling in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*,1–43.

American Education Research Association [AERA] Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448–452.

American Statistical Association [ASA]. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education. Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*, 37–66.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77–86.

Braun, H. I. (2004). *Value-added modeling: What does due diligence require?* Princeton, NJ: Educational Testing Service.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Doran, H., & Izumi, L. T. (2004). *Putting education to the test: A value-added model for California.* San Francisco, CA: Pacific Research Institute.

Doran, H. C., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 80–104). Maple Grove, MN: JAM Press.

Doran, H. C., & Fleischman, S. (2005). Challenges of value-added assessment. *Educational Leadership, 63*(3), 85–87.

Florida Department of Education. (2003). Florida Comprehensive Assessment Test (FCAT): Assessment and School Performance.

Green, J. L. (2010). *Estimating teacher effects using value-added models*. University of Nebraska at Lincoln, Department of Statistics: Dissertations and Theses.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted*? Education Finance and Policy, 10*(1), 117–156.

Guzman, N. L. (2016). Review of rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability. *Education Review//Reseñas Educativas*, 23.

Hanushek, E. A. (1972). *Education and Race: An analysis of the educational production process.* Lexington, MA: Lexington Books.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review*, *21*, 1–17.

Manzi, J., San Martín, E., & Van Bellegem, S. (2014). School system evaluation by value added analysis under endogeneity. *Psychometrika, 79*, 130–153.

McCaffrey, D., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. Washington, DC: RAND.

Murnane, R. J. (1975). *The impact of school resources on the learning of children.* Cambridge, MA: Ballinger Publishing.

Olson, L. (2004, November 16). "Value added" models gain in popularity. *Education Week.* Retrieved from http://www.edweek.org/ew/articles /2004/11/17/12value.h24.html

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*, 121–129.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1–17.

Raudenbush, S., & Bryk, A. (1988-89). Methodological advances in studying effects of schools and classrooms on student learning. In E. Z. Roth (Ed.), *Review of research in education* (pp. 423–475). Washington, DC: American Educational Research Association.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

Sanders, W. L. (2000). Annual CREATE Jason Millman Memorial Lecture: Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*, 329–339.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311.

Sanders, W. L., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*, 247–256.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Educational Assessment System (TVAAS): A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Sanders, W. L., Saxton, A., Schneider, J., Dearden, B., Wright, S. P., & Horn, S. (2002). *Effects of building change on indicators of student achievement growth: Tennessee Value-Added Assessment System.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Washington, DC: Institute for Education Sciences.

Stewart, B. E. (2006). *Value-added modeling: The challenge of measuring educational outcomes.* New York, NY: Carnegie Corporation of New York.

Tekwe, C. D., Carter, R. L., Ma, C-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*, 11–36.

Wainer, H. (2004). Introduction to a special issue of the journal of educational and behavioral statistics on value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 1–3.

Wiley, E. W. (2006). *A practitioner's guide to value-added assessment*. Retrieved from http://nepc.colorado.edu/files/Wiley_APractitionersGuide.pdf

GENİŞ ÖZET

Giriş

Son yıllarda, okulların etkililiği ve hesap verebilirliği konularına ilgide dünya çapında bir artış gözlenmektedir. Bu konulardaki ilk uygulamalar, öğrencilerin mevcut başarı durumlarını kullanmaya odaklanmıştır. Mevcut durum yaklaşımı, farklı kademedelerdeki öğrencileri tek bir zaman noktasında (genellikle dönem sonunda) karşılaştırmaya dayanmaktadır (Doran ve Izumi, 2004). Eğitimciler, bir seferlik test puanını kullanarak öğrencilerin performansı üzerindeki okul etkilerini tahmin etmenin çok doğru bir yol olmadığını düşünmektedir. Bu nedenle hesap verebilirlik sisteminde okul etkinliğini değerlendirmenin alternatif yolları aranmıştır. Bu yeni yaklaşımlar öğrencilere okulda yıl boyunca yapılan iyileştirmeler üzerine odaklanmaktadır. Araştırmacılar etkili okulları belirlemek için bireysel öğrenci başarı puanlarını zamanla beraber kullanmalarını sağlayan katma-değerli değerlendirme (KDD) fikrini geliştirmiştir. Tekwe ve diğerleri (2004)'e göre "Katma değer ifadesi bir yıldan diğerine okul ya da öğretmenin değerlendirilmesi yöntemlerini ifade eden ve daha sonra bu ölçütün bir performans değerlendirme sistemi için temel teşkil etmesinde kullanılan bir terimdir" (s.31). KDD'nin öncüleri, KDD'nin yalnızca bir yılın başarısını ölçen standart testlerden elde edilen sonuçlardan (mevcut durum yaklaşımı) daha adil ve daha doğru tahminler ürettiğini iddia etmektedir. KDD'nin birincil amacı öğretmenlerin veya okulların öğrencilerin gelişimine olan etkilerini belirlemektir (Raudenbush, 2004). KDD sistemi, okulları öğrencilerin bilgi düzeylerini nasıl arttırdıklarına göre değerlendirmeye dayanır.

Bugüne kadar, okul etkililiğinin değerlendirilmesinde basit gelişim puanlarından karmaşık karma modellere kadar değişen çeşitli alternatif modeller (Katma-Değerli Modeller; KDM) önerilmiş olmasına rağmen bu modelleri karşılaştıran sınırlı sayıda çalışma bulunmaktadır (McCaffrey vd., 2003; Tekwe ve diğerleri, 2004; Weiss, 2006). Hesap verebilirlik sistemlerinde yeni KDD yaklaşımlarını benimseyerek problemlere çözüm bulmak adına hangi modelin en etkili ve hangi modelin en kolay uygulanabilir olduğunun gösterilmesinin uygulamacılar adına faydalı olacağı düşünülmektedir. Herhangi bir KDD modeli geleneksel yöntemlerden daha üstün olmakla birlikte devletlerin hesap verebilirlik sistemlerinin (karmaşıklığından ötürü) KDM'leri kullanmadaki isteksizliği gözlenmektedir. KDM'ler açısından daha basit modellerin daha karmaşık modeller kadar etkili olduğunu gösteren çalışmalara uygulamacıların fikrini değiştirmek adına ihtiyaç duyulmaktadır.

KDM'lerin geleneksel yöntemlerden daha etkili olduğu görüşünün yanında bu modellerin ve dayandığı istatistiksel uygulamaların doğru ve güvenilir sonuçlar üretmediğini ileri süren çalışmaların olduğu da unutulmamalıdır (Guarino, Reckase ve Wooldridge, 2015; Manzi, San Martín ve Van Bellegem, 2014; Schochet ve Chiang, 2010; Wiley, 2006).

KDD kapsamında geliştirilen modeller farklı varsayımlara dayanarak okul katma değerlerini hesaplamaktadır. Örneğin, bazı modeller okula ait ve okul dışı diğer değişkenleri hesaba katmazken, bazı modeller bu faktörleri istatistiksel düzenlemelerle kontrol etmeye çalışmaktadır. Florida Comprehensive Assessment Testi'nden (FCAT) elde edilen sınav puanlarını kullanarak; bu çalışmada, okula ait ve okul dışı faktörlerin okul düzeyindeki katma-değer puanlarına etkileri araştırılmış ve KDM'lerin karmaşıklığı konusu üzerine ışık tutulmaya çalışılmıştır. Bu iki konu bağlamında uygulayıcılar ve eğitim yöneticileri için en faydalı modeli/modelleri belirlemek amacıyla en yaygın olarak kullanılan üç katma-değerli model incelenmiştir. Bu çalımada cevaplanmaya çalışılan temel soru: "Okul etkinliğinin katma-değerli değerlendirilmesi için karmaşık istatistiksel modellere gerçekten ihtiyacımız var mı, yoksa daha basit modellerle daha karmaşık modellerle olduğu kadar okul etkililiğini etkin bir şekilde değerlendirebilir miyiz?"

Yöntem

Bu çalışmada, 2003 yılında Florida eyaletinde bulunan orta okul (6-8. sınıflar) öğrencilerine ait verilerin ayrı ayrı analizleri yapılmıştır. Öğrencilerin FCAT matematik ve okuma testlerinden 2002 ve 2003 yıllarında aldıkları puanlar büyük bir bölgedeki 44 okulun katma-değerlerini tahmin etmek için analiz edilmiştir. Analizlerde sadece standart müfredatı takip eden öğrenciler kullanılmıştır; özel eğitim öğrencileri ve sınırlı İngilizce yeterlik programında iki veya daha az yıl geçiren öğrenciler de analizlerin dışında tutulmuştur. Bu çalışmada toplam 60.718 öğrenci bulunmaktadır. Yoksulluk durum bilgisi, bir öğrencinin ücretsiz öğle yemeği alıp almayacağına bağlı olarak belirlenmiştir. Diğer okul dışı değişken, etnik köken değişkeni olarak tanımlanmıştır. Bu çalışmada, okul etkililiği bağlamında üç popüler KDM (basit sabit etki modeli (SFEM) ve iki hiyerarşik doğrusal model (düzeltilmiş HLM: AHLMM ve düzeltilmemiş HLM: UHLMM)) incelenmiştir.

Bulgular ve Sonuç

Bu çalışmada kullanılan katma-değerli modellerden elde edilen okul katma değer tahminleri, modellerin farklı özelliklerinin bu tahmin değerleri ve okul etkililiğini belirlemedeki etkilerini görmek için incelenmiştir. Birincil soru, SFEM gibi daha basit modellerin okul sıralaması açısından AHLMM gibi daha karmaşık olan modeller kadar etkili olup olmadığını araştırmaktı. Önceki araştırmalar, basit ve karmaşık modellerin sonuçları arasında çok az farklılıklar olduğunu bulmuştur (Tekwe ve diğerleri, 2004). Bu çalışmadaki analizlere göre SFEM ve AHLMM arasındaki korelasyon ,55 ile ,85 arasında değişmektedir. Bu çalışmanın sonucu Tekwe ve diğerleri (2004) bulgularıyla kısmen tutarlılık göstermektedir. Basit model (SFEM) okul sıralaması açısından AHLMM ile benzer sıralamalar üretmiştir. Ayrıca, basit modellerin karmaşık modeller kadar etkili olduğunu ve bu basit modelin (SFEM) çalışmada ele alınan daha karmaşık modellerin (AHLMM ve UHLMM) yerine geçebilecekleri sonucuna varılmıştır. Uygulamacılar arasında daha basit istatistiksel modelleri kullanma isteği olduğundan, bu sonuçlar önceki araştırmalara ek olarak basit modellerin de karmaşık modeller kadar hesap verebilirlik sisteminde etkili olabileceğini göstermektedir.

Bu çalışmada okula ve öğrenciye ait değişkenlerin katma-değer tahminleri üzerine etkisi de incelenmiştir. Modeller arasında sadece AHLMM okul tahmin değerlerini etkileyebilecek bu karıştırıcı değişkenleri kontrol edebilen istatistiksel düzeltmelere sahiptir. Tekwe ve diğerleri (2004), KDD analizlerinde bu değişkenlerin modele doğrudan dahil edilmesinin veya tamamıyla göz ardı edilmesinin, okullar hakkında yanlı tahminler elde edilmesine yol açtığını belirtmektedir. Araştırmacılar bunu yapmak yerine bu değişkenleri istatistiksel olarak kontrol edebilen modellerin kullanılmasını tavsiye etmektedir. Çalışmamızda AHLMM'de bu değişkenlerin etkisini görmek için karşılaştırmalar yapılmıştır. AHLMM'nin sonuçları ile UHLMM'nin ve SFEM'nin sonuçları arasında belirgin bir farklılık bulunamamıştır. Genel olarak, bu eş değişkenlerin dâhil edilmesinin, katma değerli tahminler üzerinde büyük bir etkisi olduğu sonucuna varabiliriz. Bu sonucun aynı zamanda Tekwe ve diğerleri (2004)'ün yorumlarıyla da uyumlu olduğu görülmektedir. Sonuçlara dayanarak, öğrencilerin farklı arka planlara sahip olduğu durumlarda diğer VAM'lara nazaran AHLMM'nin tercih edilmesini tavsiye edebiliriz.

Çalışmamızın sınırlılıklarından birisi de okul değerlendirmesinde sıklıkla kullanılan LMEM'nin kullanmış olduğumuz veri yapısından dolayı çalışmaya dâhil edilmemiş olmasıdır. LMEM, okul, konu ve yıl açısından çoklu durumları dikkate alan güçlü bir modeldir. İki yıllık veri ve istikrarlı öğrenciler nedeniyle LMEM'nin gerçek etkisini çalışmamızda göremeyeceğimiz düşüncesiyle analizler arasına eklenmemiştir. Çok değişkenli yöntemin okulun etkinliği üzerindeki etkisini görmek için daha fazla araştırmanın farklı veriler kullanarak yapılması önerilir.

Sonuç olarak, hesap verebilirlik sistemini şekillendirmede KDM'lerin önemli bir rolü olduğu bu çalışmada gösterilmeye çalışılmıştır. Bu çalışmada elde edilen bulgular Florida Eyaletinde uygulanan FCAT sınavından elde edildiği için, çalışmanın bulgularının diğer eyaletlere ya da ülkelere genellenip genellenemeyeceği kesin olmamakla beraber alan yazında KDD modellerinin kullanıma dair ek kanıtlar sunduğu açıktır. Ayrıca bu çalışmada katma-değerli değerlendirme yaklaşımı ve uygulanmasında kullanılan modeller tartışıldığı için çalışmanın okul etkililiği üzerine çalışan yöneticiler ve eğitimciler için faydalı olacağı düşünülmektedir. Yurt dışında birçok ülkede tercih edilen ve okul performansının değerlendirilmesinde kullanılan bu modellere ait ayrıntılı açıklamalar içeren bu çalışmanın ülkemizde bu modelleri uygulamak isteyen araştırmacılara yardımcı olacağı düşünülmektedir.

Appendices

Appendix A. Grade 6 Math Estimates

Table 5. Estimates of the School Effects Obtained from Three VAMs Based on Grade 6 Math Results

	SFEM		UHLMM		AHLMM	
Rank*	Estimate	School ID	Estimate	School ID	Estimate	School ID
1	54.126	34	47.539	25	43.963	25
2	50.883	10	45.621	13	22.978	19
3	46.729	41	41.297	19	19.502	6
4	43.380	39	39.370	34	18.071	22
5	32.629	42	29.861	41	17.040	13
6	32.055	14	28.624	6	14.723	32
7	31.476	16	27.787	10	14.085	37
8	25.660	11	23.992	14	11.225	23
9	24.598	13	22.195	11	10.468	31
10	24.186	1	21.982	8	10.446	41
11	24.036	28	21.971	42	10.196	4
12	22.312	6	19.878	12	9.751	27
13	21.740	26	19.528	37	9.287	12
14	19.308	12	17.133	39	9.253	34
15	10.485	8	9.613	29	7.079	11
16	10.254	7	9.203	36	7.000	42
17	9.985	33	8.964	4	6.872	30
18	8.621	2	7.843	28	5.885	8
19	7.804	43	6.678	22	4.198	38
20	6.766	36	5.929	24	3.468	10
21	3.580	30	3.196	26	2.105	29
22	1.377	38	1.181	38	1.489	39
23	0.695	18	0.552	5	1.349	36
24	-4.603	24	-4.034	15	1.149	14
25	-6.922	19	-6.003	31	1.092	9
26	-9.650	15	-8.645	17	-1.146	24
27	-9.718	29	-8.779	32	-3.735	17
28	-10.151	25	-9.158	23	-4.549	5
29	-10.366	31	-9.277	2	-4.838	3
30	-13.212	23	-11.659	7	-5.679	28
31	-13.489	37	-12.163	40	-6.250	20
32	-18.218	20	-16.900	1	-6.839	18
33	-19.810	40	-17.065	27	-7.952	15
34	-20.228	17	-18.407	9	-9.202	44
35	-21.194	32	-19.345	20	-9.377	7
36	-21.681	35	-19.563	30	-10.226	26
37	-24.274	4	-22.656	16	-12.559	35
38	-32.380	5	-28.937	3	-14.295	40
39	-33.237	3	-29.809	21	-15.309	2
40	-34.008	22	-30.654	18	-17.635	21
41	-50.386	44	-41.325	44	-23.673	16
42	-53.935	21	-45.737	43	-26.979	1
43	-58.680	9	-49.734	33	-39.858	43
44	-	27	-50.094	35	-42.581	33

Note. Only the school rankings based SFEM estimates were presented in the table. Estimate represents fixed effect estimates for SFEM while random effects estimates are presented for UHLMM and AHLMM. SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

Appendix B. Grade 6 Reading Estimates

Table 6. Estimates of the School Effects Obtained from Three VAMs Based on Grade 6 Reading Results

	SFEM		UHLMM		AHLMM	
Rank*	Estimate	School ID	Estimate	School ID	Estimate	School ID
1	60.918	25	44.015	25	-24.438	36
2	48.315	10	35.401	10	-23.498	33
3	36.844	13	27.928	13	-16.870	43
4	27.022	34	21.091	34	-14.180	1
5	24.526	39	19.427	14	-13.321	21
6	23.067	14	18.243	28	-11.845	42
7	22.844	28	18.239	39	-10.927	35
8	22.459	8	17.915	8	-10.611	38
9	17.803	26	13.995	26	-9.648	24
10	15.573	19	12.195	11	-8.487	6
11	15.465	2	11.993	2	-6.880	9
12	15.173	11	11.402	19	-6.373	16
13	12.764	29	10.336	29	-5.099	4
14	11.373	27	7.840	20	-4.015	41
15	10.610	20	7.753	27	-4.011	17
16	10.071	12	7.536	12	-2.407	18
17	8.149	5	6.051	5	-1.944	40
18	6.386	32	4.879	32	-1.771	15
19	5.408	40	4.064	40	-0.895	5
20	3.358	23	2.388	23	-0.093	37
21	3.189	15	2.180	15	0.115	7
22	2.815	31	1.899	31	0.587	12
23	1.086	7	0.756	7	2.018	3
24	0.765	37	0.560	37	2.194	22
25	-1.526	41	-1.245	41	2.909	19
26	-6.618	22	-4.619	22	3.048	44
27	-7.420	16	-6.164	16	3.343	26
28	-9.302	17	-6.806	17	3.457	29
29	-12.984	4	-9.036	44	3.732	2
30	-13.007	3	-9.660	24	3.736	11
31	-13.095	24	-9.730	3	4.714	30
32	-14.541	1	-10.055	4	5.163	14
33	-16.027	30	-11.841	1	6.266	8
34	-17.307	6	-12.910	30	6.417	34
35	-19.570	42	-13.058	6	6.687	28
36	-22.921	18	-14.798	42	7.251	39
37	-24.321	43	-18.471	18	8.767	20
38	-25.337	38	-18.894	43	9.931	31
39	-26.937	21	-19.247	21	10.142	23
40	-29.138	9	-19.764	38	11.042	13
41	-35.153	33	-22.641	9	13.489	27
42	-44.516	36	-28.362	33	14.161	32
43	-54.027	35	-34.359	36	15.269	10
44	-	-	-36.434	35	32.868	25

Note. Only the school rankings based SFEM estimates were presented in the table. Estimate represents fixed effect estimates for SFEM while random effects estimates are presented for UHLMM and AHLMM. SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

Appendix C. SAS Codes Used for Model Estimations

```
*/ SAS Code for Model1 (SFEM)*/;
proc glm data=GRADE6;
model cahangem = S1 - S43/solution; run;
```

```
*/ SAS Code for Model2 (UHLMM)*/;
proc mixed data=GRADE6;
class student;
model changem =;
random intercept / type = un sub = school solution;
repeated /type = un sub = student; run;
```

```
*/ SAS Code for Model3 (AHLMM)*/;
proc mixed data=GRADE6;
class student min pov;
model changem = Z1M Z2 V1 min pov;
random intercept/type= un sub = school solution;
repeated/type = un sub = student; run;
```

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);321-343

The Number of Response Categories and the Reverse Scored Item Problem in Likert-Type Scales: A Study with the Rasch Model*

Likert Tipi Ölçeklerde Olumsuz Madde ve Kategori Sayısı Sorunu: Rasch Modeli ile Bir İnceleme

Mustafa İLHAN** Neşe GÜLER***

Abstract

This study aims to address reverse scored item and the number of response categories problems in Likert-type scales. The Fear of Negative Evaluation Scale (FNES) and the Oxford Happiness Questionnaire (OHQ) were used as data collection tools. The data of the study were analyzed according to the Rasch model. It was found that the observed and expected test characteristic curves were largely overlapped, each of the three rating scales worked effectively, and the differences between response categories could be distinguished successfully by the participants in straightforward items. On the other hand, it was determined that there were significant differences between the observed and expected test characteristic curves in reverse scored items. According to the results the participants could not distinguish the response categories of the reverse scored items at three, five and seven-point rating versions of both scales. Hence, the reverse scored items were removed from the data file, and the analysis was repeated. The results revealed that item discrimination, reliability coefficients for person facet, separation ratios and Chi square values calculated for the facets of person and items were higher in five-pointed rating compared to three and seven pointed rating. Based on these results it can be said that the scale categories in reverse scored items could not be discriminated by responders at all type of rating, and that reverse scored items did not measure the same latent structure as straightforward items did.

Key Words: Likert type scale, reverse scored item, number of response categories, Rasch model

Öz

Bu araştırmada Likert tipi ölçeklerde olumsuz madde ve kategori sayısı sorununun ele alınması amaçlanmıştır. Çalışmada veri toplama aracı olarak Olumsuz Değerlendirilme Korkusu Ölçeği (ODKÖ) ile Oxford Mutluluk Ölçeği (OMÖ) kullanılmıştır. Araştırma kapsamında toplanan veriler Rasch modeline göre analiz edilmiştir. Analiz sonucunda; ODKÖ ile OMÖ'deki olumlu maddelerde gözlenen ve beklenen test karakteristik eğrilerinin büyük ölçüde örtüştüğü, her üç kategori sayısının da etkin bir biçimde çalıştığı ve ölçek kategorileri arasındaki farkların katılımcılar tarafından başarılı bir biçimde ayırt edildiği belirlenmiştir. Öte yandan olumsuz maddelerde gözlenen ile beklenen test karakteristik eğrileri arasında önemli farklılıklar olduğu saptanmıştır. Üç, beş ve yedili derecelendirmeden hangisi kullanılırsa kullanılsın, ODKÖ ile OMÖ'deki olumsuz maddelerde kategorilerin katılımcılar tarafından ayırt edilemediği tespit edilmiştir. Bu tespitin ardından olumsuz maddeler veri dosyasından çıkarılarak analiz tekrarlanmıştır. Elde edilen bulgular; madde ayırt ediciliği, birey yüzeyine ilişkin güvenirlik katsayısı ile birey ve madde yüzeyleri için hesaplanan ayırma oranı ve Ki Kare değerlerinin beşli derecelemede üçlü ve yedili derecelemeye göre daha yüksek olduğunu göstermiştir. Bu bulgular, üçlü, beşli ya da yedili derecelemeden hangisi kullanılırsa kullanılsın olumsuz maddelerde ölçek kategorilerinin cevaplayıcılar tarafından ayırt edilemediğine ve olumsuz maddelerin olumlu maddelerle aynı örtük yapıyı ölçmediğine işaret etmektedir.

Anahtar Kelimeler: Likert tipi ölçek, olumsuz madde, kategori sayısı, Rasch modeli

* It was orally presented at the 5th Congress on Measurement and Evaluation in Education and Psychology.

** Asst. Prof., Dicle University, Faculty of Education, Diyarbakir-Turkey. e-posta: mustafailhan21@gmail.com, ORCID ID: orcid.org/0000-0003-1804-002X

*** Assoc. Prof., Sakarya University, Faculty of Education, Sakarya-Turkey. e-posta: gnguler@gmail.com, ORCID ID: orcid.org/0000-0002-2836-3132

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 3, Sonbahar 2017, 321-343.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 3, Autumn 2017, 321-343.
Received: 12.06.2017
Accepted:14.09.2017
DOI: 10.21031/epod.321057

INTRODUCTION

Likert type scales were introduced to the literature by Rensis Likert in 1932 (Likert, 1932). A number of statements are presented to participants in such scales and they state the extent to which they agree with the statement on a continuum ranging between *strongly agree* and *strongly disagree* or between *very appropriate to me* and *not appropriate to me at all* (Erkuş, 2003). The development and implementation of Likert type scales are easier than other measurement tool (Ahlawat, 1985; Tezbaşaran, 1997; Tavşancıl, 2010). Therefore these scales are frequently used in research in social sciences, psychology and educational sciences (Adelson & McCoach, 2010; Chang, 1994). Due to their common use, a great number of studies were performed in relation to determining how changes in the format of Likert type scales affect the psychometric properties of measurements. How reverse scored items used in Likert type scales influence measurement results, and what the most appropriate number of response categories is in such scales are the two basic issues considered in the studies (Preston & Colman, 2000).

The Problem of Reverse Scored Item

One of the most fundamental problems in Likert type scales is about how useful the reverse scored items are in such scales, and about how validity and reliability are influenced by them. Reverse scored items are also known as negative items, and the high scores received from these items indicate that participants have the measured psychological structure at low levels (Chiorri, Anselmi & Robusto, 2009). Developing the scale in a manner as to include reverse scored as well as straightforward items is a commonly preferred practice in order to prevent response sets based on stereotyped judgement and to reduce bias in responses such as affirmation or agreement, and social desirability (Hooper, Arora, Martin & Mulis, 2013; Van Sonderen, Sanderman & Coyne, 2013). However, there can also be disadvantages in using straightforward and reverse scored items together in the scale (Zhang, Noor & Savalei, 2016). Hence, DeVellis (2003) calls attention to the fact that there can be a cost in including reverse scored items in the scale, and says that those items can cause confusion in responders. It is possible to come across with empirical studies overlapping with this view of DeVellis (2003) in the literature. For instance, Schrieheim and Hill (2003) conclude that reverse scored items which are used to raise validity by reducing acquiescence response bias cause decrease in validity on the contrary. Chamberlain and Cummings (1984) compared the reliability coefficients of the form containing straightforward and reverse scored items with those of the form containing only straightforward items. In consequence, they found that the reliability coefficient for the form containing only straightforward items was higher. Hooper et al (2013) found that reverse scored items made measurement models complicated and that they caused inclusion of variance irrelevant to the measured structure in results. Locker, Jokovic and Allison (2013) found that straightforward and reverse scored items were in different factors, and this result was considered as evidence for the fact that reverse scored items did not measure the same latent structure as straightforward items did. Conrad et al. (2004) analysed the restrictions of reverse scored items through Rasch analysis reaching similar conclusions, and they found that reverse scored items in the scale caused a decrease in model-data fit. In contrast to these studies pointing to the fact that reverse scored items did not work well and that they could threaten validity, Bergstrom and Luriz (1998) found that straightforward and reverse scored items measured the same structure, and that using these two types of items together was unobjectionable. Thus, it may be stated that there is no consistency between research findings on how reverse scored items influence validity and reliability.

Number of Response Categories Problem

The most frequently preferred number of response categories in Likert type scales is five-pointed rating recommended by Likert (1932) (Lozano, García-Cueto & Muñiz, 2008). However, a review of literature demonstrates that differing number of response categories can be used and that the issue of the most appropriate number is controversial. How the number of categories in the scale affects the averages and variance values of measurements, the distribution of the data, and the skewness and kurtosis coefficients (Dawes, 2008), and what categories participants preferred more (Preston & Colman, 2000) can be the subject matter of such controversy. In addition to that, the probable effects of the number of response categories on the validity and reliability of measurements are also the focus of those discussions (Turan, Şimşek & Aslan, 2015).

On examining the studies concerning how the number of response categories in a scale affects reliability, it was found that the majority of them (Aiken; 1983; Birkett, 1986; Chang, 1994; Cicchetti Shoinralter & Tyrer, 1985; Halpin, Halpin & Arbet, 1994; Jenkins & Taber, 1977; Lissitz & Green, 1975) were based on internal consistency reliability, and that relatively small portion of them (Boote, 1981; Oaster, 1989; Ramsay, 1973; Weng, 2004) were based on the effects of test-retest reliability. Those studies obtained inconsistent findings on the effects of response categories on both internal consistency reliability and test-retest reliability. Studies performed by Aiken (1983), Leung (2011), and Qasem, Almoshigah and Gupta (2014), for instance, found that the number of response categories in a scale did not have any effects on internal consistency reliability. Weng (2004), Lozano, Garcia-Cueto and Muniz (2008) and Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol and Coffman (2009), on the other hand, concluded that internal consistency reliability tended to rise in parallel to the increase in the number of response categories. While Matell and Jacoby (1971) found that the number of response categories did not affect test-retest reliability, Oaster (1989) and Weng (2004) reported that test-retest reliability rose with an increase in the number of response categories.

An examination of the studies analysing the correlations between the number of response categories in a scale and validity demonstrated that some of the studies tested the effects of the number of response categories on construct validity, and that some others sought answers to the question of whether or not criterion-based reliability differed according to the number of response categories used. Varied results were obtained in relation to the effects of the number of response categories on validity in the studies conducted. For example, Comrey and Montang (1982), and King, King and Klockars (1983) found that the rate of total variance explained and factor loads was higher and factor structure was more clear in seven-pointed rating than in two-pointing rating. Lozano, García-Cueto and Muñiz (2008) also found that the rate of variances explained in the factor analysis rose as the number of response categories in the scale increased; and the result was interpreted as that the increase in the number of response categories influenced validity in positive ways. In a similar vein, Tarkan (2015) concluded that factorial validity increased as the number of response categories in a scale increased. In contrast to these studies, Kim (1998) found in the study comparing three-pointed, five-pointed, seven-pointed and nine-pointed rating in terms of validity and reliability that validity was the lowest in three-pointed rating, medium in seven-pointed rating, and it was higher in five and nine-pointed rating. The study conducted by Maydeu-Olivares (2009) found that model-data fit decreased as the number of response categories in a scale increased.

There is no overlap between the findings for the effects on criterion-based validity as in the findings concerning the effects of the number of response categories on construct validity. In the study performed by Chang (1994) where four-pointed and six-pointed ratings were compared psychometrically, it was found that the number of response categories did not have any effects on criterion-based validity. In a similar way, Qasem, Almoshigah and Gupta (2014) also concluded that there were no significant differences between the criterion-based validity coefficients of the scales with two, three and five-pointing rating. Loken, Pirie, Virnig, Hinkle and Salmon (1987), on the other hand, point out that criterion-based validity is influenced by the number of response categories and that the criterion-based validity coefficients obtained from 11-pointed rating are higher than those calculated from three or four-pointed rating. However, Preston and Colman (2000) found that the criterion-based validity coefficients of the scales using two, three and four-pointed rating were lower, and the criterion-based validity coefficients of the scales with five or more categories were higher. Besides, it was also found that there were no statistically significant differences between the criterion-based validity coefficients of the scales having differing numbers of response categories.

The Purpose and Significance of the Study

This study has basically two purposes. First, it aims to determine the extent to which reverse scored items in Likert type scales are functional. In accordance with this aim, *i)* whether or not scale categories functioned in the same way in straightforward and reverse scored items was evaluated; *ii)* efforts were made to determine whether or not those items measured the same latent structure by comparing the test characteristic curves for those straightforward and reverse scored items. The operations mentioned were performed separately for Likert type scales having three, five and seven-pointed rating; and thus it was checked whether or not the functioning of the reverse scored items was influenced by the number of response categories in scales. Secondly, the study aims to exhibit the effects of the number of response categories used in Likert type scales on the psychometric properties of measurements. In line with this purpose, Likert type scales having three, five and seven-pointed rating were compared in terms of reliability and model-data fit. In this way, validity was not ignored while the effects of the number of response categories on reliability were being analysed. This is quite important to make study results more meaningful because, as it is also pointed out by Cronbach (1950), it is worthless to increase reliability on its own and validity should also be taken into consideration in order to be able state that a certain number of response categories raising reliability is appropriate.

This study differs from the previous studies in the literature in several aspects. Therefore, it is predicted that the study will contribute significantly to the relevant literature. Firstly, the number of response categories and reverse scored items are considered separately in the studies available in the literature. This situation leaves the question of whether the differences in the number of response categories in Likert type scales have the same effects on straightforward and reverse scored items unanswered. In other words, while the most appropriate number of response categories for Likert type scales was investigated in the studies in the literature, general evaluations were made based on the items, but the effects of the number of response categories on straightforward and reverse scored items were not tested separately. Because this current study considers the number of response categories and the problem of reverse scored items together, it will be possible here to reveal how the differences in the number of response categories influence measurement results separately for straightforward and reverse scored items. Hence, this study differs from other studies in the literature in this respect.

Secondly, all of the studies in the literature aiming to determine the most appropriate number of response categories in Likert type scales and the way reverse scored items used in those scales affect the psychometric properties of measurements were conducted in other cultures. No studies are available in relation to Turkish culture. Yet, cultural properties have significant effects on the responses given to Likert type items. The significant effects were shown in many studies in the literature. For instance, Bachman and O'Malley (1984) found that there were differences between sets of responses given by white and black individuals in the USA, and that the likelihood of the blacks to use the extreme points in Likert type scales was higher than that of whites. Stening and Everett (1984), however, found that the likelihood of Japanese people to use the middle points was higher than that of American or British people in answering the Likert type scales. Huri and Triandis (1989) compared the sets of responses given by Spanish and non-Spanish participants through a Likert type scale of five and ten-point rating. Accordingly, it was shown that Spanish participants used the extreme points of the scales more frequently than the others. In addition to that, it was also found that the sets of extreme responses given by Spanish participants decreased on using 10-pointed rating, and that the answers given by non-Spanish participants were not influenced by the number of response categories in the scale. In a study conducted by Hofsteder (1998), it was found that the participants living in masculine cultures used extreme points more in answering the scale items, but that the participants coming from dominantly feminine cultures tended to use answers in the middle points. Lee, Jones, Mineyama and Zhang (2002) found that the likelihood of the Japanese and the Chinese to choose the middle points in Likert type scales was higher than that of Americans. Whether individualism or collectivism has priority in cultures is another issue influential in responses to Likert type scales. Johnson, Kulesa, Cho and Shavitt (2005) found that the probability of giving responses corresponding to the extreme points in the scale was higher in cultures where individualism is stressed. In cultures where collectivism is dominant, on the other hand, it was more probable to give responses in the middle points of the scale. Hooper et al. (2013) studied how reverse scored items in mathematics self-efficacy scale which was applied in Trends in International Mathematics and Science Study (TIMSS) worked. Accordingly, it was found that the effects of reverse scored items on model-data fit differed from one country to another. On considering these studies indicating that cultural factors influence the responses to Likert type scales, it may be said that it is difficult to generalise the conclusions reached in studies conducted in different cultures into Turkish culture. In this sense, it is hoped that such a study to be performed would contribute to literature.

Thirdly, although there are numerous studies in the literature trying to determine whether or not the psychometric properties of Likert type scale differ according to the number of response categories in a scale, they dominantly use methods based on Classical Test Theory (CTT). For this reason, there can be some points in which the above mentioned studies are inadequate in answering the question of how the psychometric properties of measurements are influenced by the number of response categories used in a scale. One such point is that CTT-based reliability calculations in Likert type scales are restricted to item reliability. According to CTT, variability observed between individuals all stems from the personal differences of participants. Therefore, a reliability coefficient is not reported for individuals in CTT (Taşdelen, Güler & Kaya Uyanık, 2015). In Rasch model, however, individuals in addition to items are considered as the sources of error. Thus, reliability coefficients are calculated for both items and persons in the Rasch analysis. In this context, using the Rasch model in this study will enable us to determine the effects of the number of response categories on reliability for both items and persons (Güler, İlhan, Güneyli & Demir). Another point in which CTT is inadequate in determining whether or not the adopted number of response categories is appropriate is that it does not inform one of category statistics. To put it in more clear terms, it is impossible in CTT to make an evaluation on whether or not participants can distinguish between the sequential points of scale categories. In Rasch analysis, on the other hand, a table of category statistics is reported, and it is possible to make inferences about how well participants can distinguish between scale categories by analysing the table.

Finally, it is evident that the studies in the literature investigating the problem of reverse scored items in Likert type scales use one single tool of measurement. This current study, however, employs two different scales. In the first scale to be used in this study (Happiness Scale), high scores indicate positive properties for participants whereas in the other scale (The Fear of Negative Evaluation Scale) high scores represent negative properties for participants. Thus, it will be possible in this study to determine whether or not reverse scored items function in the same way in scales where high scores represent desired properties (such as self-respect, self-efficacy and job satisfaction) and in measurement tools where high scores represent undesired properties (such as anxiety, burnout and stress). It is thought that the study is also original in this respect and that it will contribute to the literature.

METHOD

Study Group

The study was conducted with two different groups which were composed of 312 university students in total. The first group consisted of 197 participants, 112 (56.90%) of whom were female and 85 (43.10%) were male. The participants' ages ranged between 17 and 34 in this group, with average age of 21.66. The second group consisted of 115 participants, 64 (55.70 %) of whom were female and 51 (44.30 %) of whom were male. The ages in this group ranged between 18 and 32, with average age of 21.72.

Data Collection Tools

The Fear of Negative Evaluation Scale (FNES) and The Oxford Happiness Questionnaire-Short Form (OHQ-S) were used as the tools of data collection.

The Fear of Negative Evaluation Scale (FNES)

The FNES, which was developed by Leary (1983), was adapted into Turkish by Çetin, Doğan and Sapmaz (2010). The original form of the scale, which was in five-pointed Likert type, contained 12 items. The Turkish version of the scale, however, it was found that the discrimination index of item four in the original form was negative. Therefore, the item was removed from the scale. Through explanatory factor analysis (EFA) performed with the remaining 11 items, a one-factor structure explaining 40.19% of the total variance was obtained. In consequence of confirmatory factor analysis (CFA), the fit indices for the one-factor model were found as: RMSEA=.062, NFI=.96, CFI=.98, IFI=.98, RFI=.95, GFI=.95 and AGFI=.92. It was determined that factor loads ranged between .44 and .78 in EFA, and between .37 and .74 in CFA. Internal consistency, split-half reliability, and test-retest reliability coefficients calculated for FNES were found to be .84, .83 and .82 respectively. Eight of the 11 items in the Turkish version of FNES were straightforward items containing statements for worries about fear of negative evaluation. The remaining three items were reverse scored items stating that there were no worries about fear of negative evaluation.

The Oxford Happiness Questionnaire-Short Form (OHQ-S)

OHQ-S was developed by Hills and Argyle (2002), and was adapted into Turkish by Doğan and Akıncı Çötok (2011). The scale is in six-pointed Likert type. It has eight items in its original form. However, item four in its Turkish version was found to have low discrimination index (.17). Thus, the item was removed from the scale, and validity and reliability studies were conducted with the remaining seven items. Following the EFA, a one-factor structure was obtained, as in the original form of the OHQ-S. It was found in this one-factor structure that the rate of explained variance was 39.74%, and that the factor loads of the items ranged between .53 and .72. The findings obtained in CFA showed that the one-factor structure of the Turkish version of the OHQ-S had adequate fit indices [χ^2/sd=2.77, RMSEA=.074, AGFI=.93, GFI=.97, NFI=.92, CFI=.95, IFI=.95 RMR=.044]. In consequence of reliability study, the internal consistency coefficient for the scale was found as .74, and test-retest reliability as .85. Five of the seven items in the Turkish form of OHQ-S were straightforward items indicating happiness whereas the remaining two were reverse scored items containing statements of unhappiness.

Data Collection

The data were collected in the spring semester of 2015-16 academic year. The data collection tools were administered to students in the classroom setting. Prior to the application, the participants were informed of the purpose of the research, and only volunteers were participated to the study. The FNES was applied to the first group, and the OHQ-S was applied to the second group in three, five and seven-pointed rating types at intervals of one week. Only extreme points were labelled in all three types of rating (*strongly disagree* → *strongly agree*), and the points between the two extremes were not labelled. The thought that a clear labelling cannot be made in such an approach as in three, five and seven-pointed rating (Şeker & Gençdoğan, 2006; Østerås et al., 2008) was influential in adopting such an approach. That is to say, differences can be observed in measurement results in Likert type scales depending on how clearly the points of a scale is labelled (Wyatt & Meyers, 1987). Therefore, it will not be possible to determine whether the findings are the results of differences in the number of response categories or of uncertainty of labelling in relation to the categories when labelling is used for all response categories in three, five and seven-pointed rating. Setting out from this fact, only extreme points were labelled in all three type of scales (in three, five and seven pointed rating).

Data Analysis

The obtained data were analysed through the Rasch model by using the FACETS package programme. Rasch model is a one-parameter model placed under the roof of item response theory (Baker, 2001). Each source of variability capable of influencing measurement results is called a facet in this model (Sudweeks, Reeveb & Bradshawc, 2004). Rasch model can be assessed under the titles of *two-facet* or *many-facet* according to the number of facets it contains. In the two-facet model, there are two sources of variability capable of influencing measurement results (Linacre, 2014) - namely, items and persons. In many-facet model, however, in addition to items and persons, there are also other sources of variability such as raters, or demographic properties for persons which can influence measurement results (Knoch & McNamara, 2015). Sources of variability capable of influencing measurement results are restricted to items and persons in this study. Therefore, two-facet Rasch model was used in the analysis of the data collected in this study. The Rasch analysis was carried out according to rating scale model and Joint Maximum Likelihood Estimation Method (Unconditional Maximum Likelihood Estimation-UCON). Rasch analysis outputs are composed of many tables and graphs such as category statistics, test characteristic curves, and measurement reports for the facets of item and person. The tables and graphs were analysed in accordance with the purpose of this study, and the analysis outputs on which each sub-purpose was based were shown in Table 1.

Table 1. Statistical Indicators Considered for each Sub-purpose of the Study

Sub-purposes	Statistical Indicators to be Considered	
To determine whether or not scale categories function in the same way in straightforward and reverse scored items	Table of Category Statistics	The statistical indicators were analysed for Likert type scales having three, five and seven-pointed rating separately to test the effects of the number of response categories on the functioning of reverse scored items.
To find whether or not straightforward and reverse scored items measure the same latent structure.	Test Characteristics Curve	
To find determine the effects of response categories on reliability	Reliability coefficients, separation ratio and Chi-square for the facets of item and person	
To demonstrate the effects of the number of response categories on validity.	Infit and outfit statistics showing the model-data fit	

RESULTS

In this part of the results of the study are presented. First, category statistics table was analysed so as to determine how actively three, five and seven-pointed rating worked. The category statistics for the straightforward items in FNES and OHQ-S are shown in Table 2, and the category statistics for reverse scored items are shown in Table 3.

Table 2. The Category Statistics for the Straightforward Items in FNES and OHQ-S

			FNES				OHQ-S			
		Category	Frequency and %	Avge Measure	Expected Measure	Outfit	Frequency and %	Avge Measure	Expected Measure	Outfit
Scaling	Three	1	848 (54%)	-1.44	-1.32	.9	81 (14%)	-.56	-.32	.8
		2	563 (36%)	-.70	-.82	.8	310 (54%)	.36	.40	.7
		3	165 (10%)	-.10	-.29	.7	184 (32%)	1.33	1.17	.9
	Five	1	465 (30%)	-.94	-.82	.8	60 (10%)	-.44	-.36	.8
		2	492 (31%)	-.65	-.64	.7	60 (10%)	-.20	-.06	.6
		3	237 (15%)	-.46	-.48	.8	146 (25%)	.14	.20	.6
		4	312 (20%)	-.16	-.31	.6	168 (29%)	.52	.48	.6
		5	70 (4%)	.07	-.12	.8	141 (25%)	.93	.82	.9
	Seven	1	637 (40%)	-.54	-.49	.9	64 (11%)	-.29	-.23	.9
		2	286 (18%)	-.49	-.42	.6	48 (8%)	-.18	-.11	.7
		3	204 (13%)	-.30	-.35	.5	57 (10%)	-.06	-.01	.8
		4	167 (11%)	-.23	-.29	.6	113 (20%)	.09	.09	.7
		5	113 (7%)	-.11	-.23	.5	101 (18%)	.23	.20	.6
		6	59 (4%)	-.07	-.17	.7	75 (13%)	.34	.32	.8
		7	110 (7%)	-.06	-.11	.8	117 (20%)	.50	.45	.9

Making at least 10 observations in each category of the scale (for instance in each of the categories 1, 2 and 3) is the first assumption to meet in order to be able to say that rating adopted in the scale works actively (Linacre, 2014). According to Table 2, there are at least 10 observations for the straightforward items in FNES and OHQ-S for each category of three, five and seven-pointed rating. The second assumption to meet is that average measurements increase monotonously (Linacre, 2014). According to the Table, the average measurements in all three, five and seven-pointed rating increase in parallel to the scale categories. In other words, there is a continuous increase in three-pointed rating as moving from category 1 to category 3, in five-pointed rating as moving from category 1 to category 5, and in seven-pointed rating as moving from category 1 to category 7. The fact that outfit statistics are within the interval of .5 and 1.5 indicates that rating on which the scale is based works well (Linacre, 2014). According to Table 2, the outfit statistics are in the .5 and 1.5 interval in all three types of rating. These findings mean that all assumptions are met to be able to say that the rating used in the scale works actively. Thus, it may be said that all three types of rating work properly with straightforward items in FNES and in OHQ-S. Having found this, the data in Table 3 were analysed so as to determine how the scale categories worked with reverse scored items.

Table 3. The Category Statistics for the Reverse Scored Items in FNES and OHQ-S

			FNES				OHQ-S			
		Category	Frequency and %	Avge Measure	Expected Measure	Outfit	Frequency and %	Avge Measure	Expected Measure	Outfit
Scaling	Three	1	155 (26%)	.12	-.35	1.8	106 (46%)	-1.14	-1.46	1.4
		2	264 (45%)	-.07*	.04	1.1	107 (47%)	-1.07	-.90	1.4
		3	172 (29%)	.17	.49	1.3	17 (7%)	-1.31*	-.42	1.9
	Five	1	47 (8%)	.23	-.14	1.8	89 (39%)	-.62	-.90	1.5
		2	189 (32%)	.24	.02	1.8	61 (27%)	-.76*	-.66	1.4
		3	104 (18%)	.28	.16	1.8	45 (20%)	-.74	-.46	2.4
		4	161 (27%)	.14*	.31	1.5	21 (9%)	-.24	-.26	1.1
		5	90 (15%)	-.03*	.47	2.0	14 (6%)	-.59*	-.06	2.6
	Seven	1	106 (18%)	.03	-.14	2.0	88 (38%)	-.45	-.60	1.3
		2	95 (16%)	-.08*	-.07	1.3	45 (20%)	-.51*	-.51	1.2
		3	84 (17%)	.07	-.01	1.5	34 (15%)	-.57*	-.42	2.2
		4	76 (13%)	.14	.05	1.0	26 (11%)	-.49	-.34	1.9
		5	41 (7%)	.18	.11	.6	20 (9%)	-.36	-.26	1.8
		6	50 (8%)	.06*	.16	1.5	6 (3%)	-.12	-.19	.7
		7	139 (24%)	.00*	.21	1.7	11 (5%)	-.43*	-.12	2.3

The symbol (*) in the table shows that the assumption that average measurements increase in parallel to the scale categories was violated.

According to Table 3, the assumption that there should be at least 10 observations in each scale category in three, five and seven-pointed rating is met. However, the assumptions that the average measurements increase along with scale categories and outfit statistics should be in the .5 and 1.5 interval are not met in any of the three, five and seven-pointed rating. Accordingly, it can be stated that the scale categories cannot be distinguished by participants in reverse scored items in FNES and OHQ-S, no matter which (three, five or seven-pointed) type of rating is used.

Following the category statistics, the test characteristics curves were analysed for FNES and OHQ-S in order to decide whether or not straightforward and reverse scored items measured the same latent structure. The test characteristic curves for straightforward items in FNES and OHQ-S are shown in Figure 1.

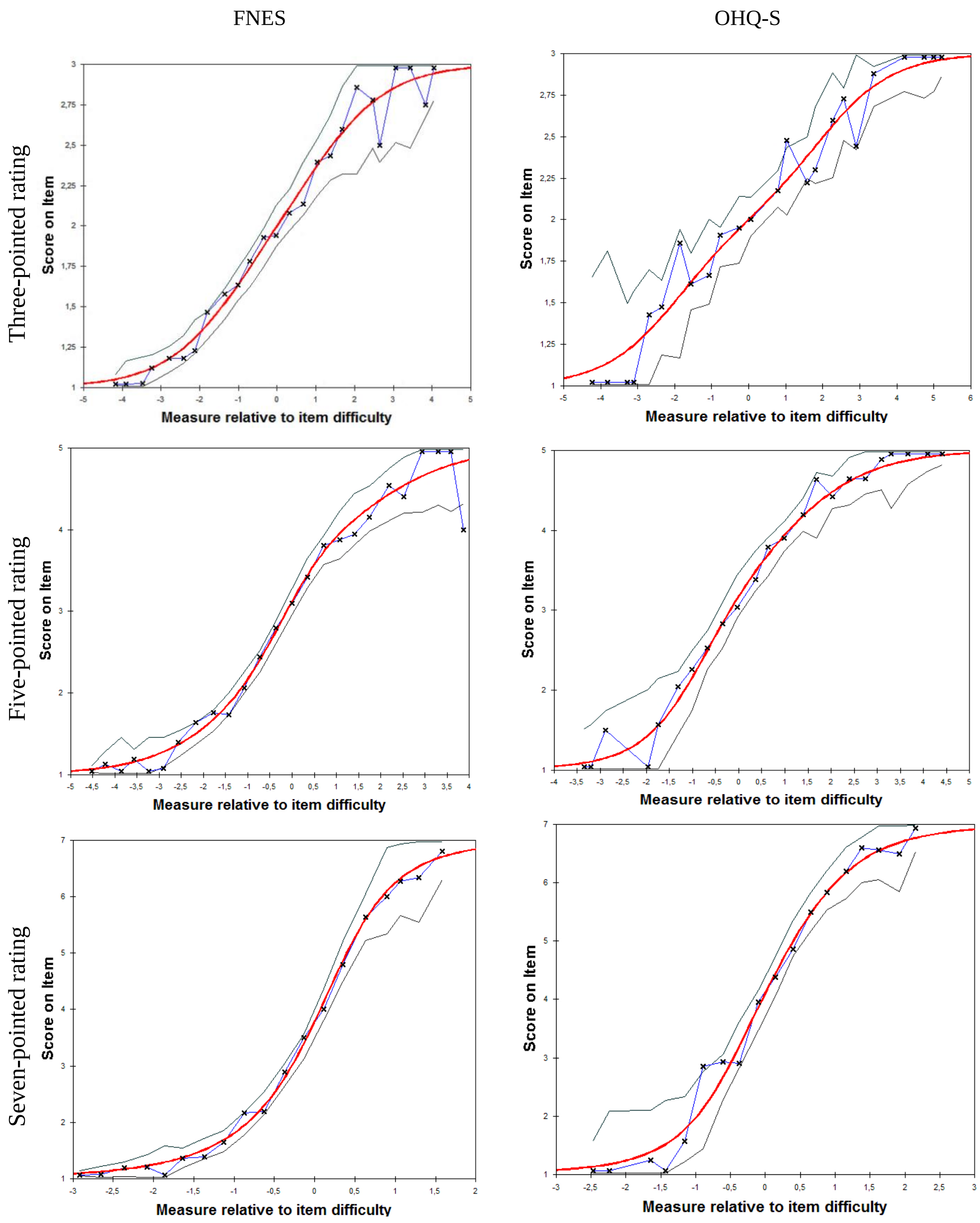


Figure 1. Test characteristic curves for straightforward items in FNES and OHQ-S.

As is clear from Figure 1, there are two lines – one of which is red and the other of which is blue on test characteristic curves. The red straight line represents the expected test characteristic curve while the blue line with crosses on it represents the observed test characteristic curve. The fact that there are no significant deviations between the expected and the observed test characteristic curves indicates model- data fit. Thus, it may be said that model-data fit is attained in all three types of rating for the straightforward items in FNES and OHQ-S. The fit shows that the straightforward items in FNES and OHQ-S can measure the latent structure which is targeted.

Test characteristic curves for the reverse scored items in FNES and OHQ-S are shown in Figure 2. According to Figure 2, there are important differences between the observed and the expected test characteristic curves for the reverse scored items in FNES and OHQ-S regardless of the type of rating. The differences show that the model-data fit is not attained in reverse scored items, and that therefore the reverse scored items in FNES and OHQ-S do not serve to measure the targeted structure. Accordingly, it may be said that the reverse scored items in FNES and OHQ-S do not measure the same latent structure as the straightforward items in FNES and OHQ-S do.

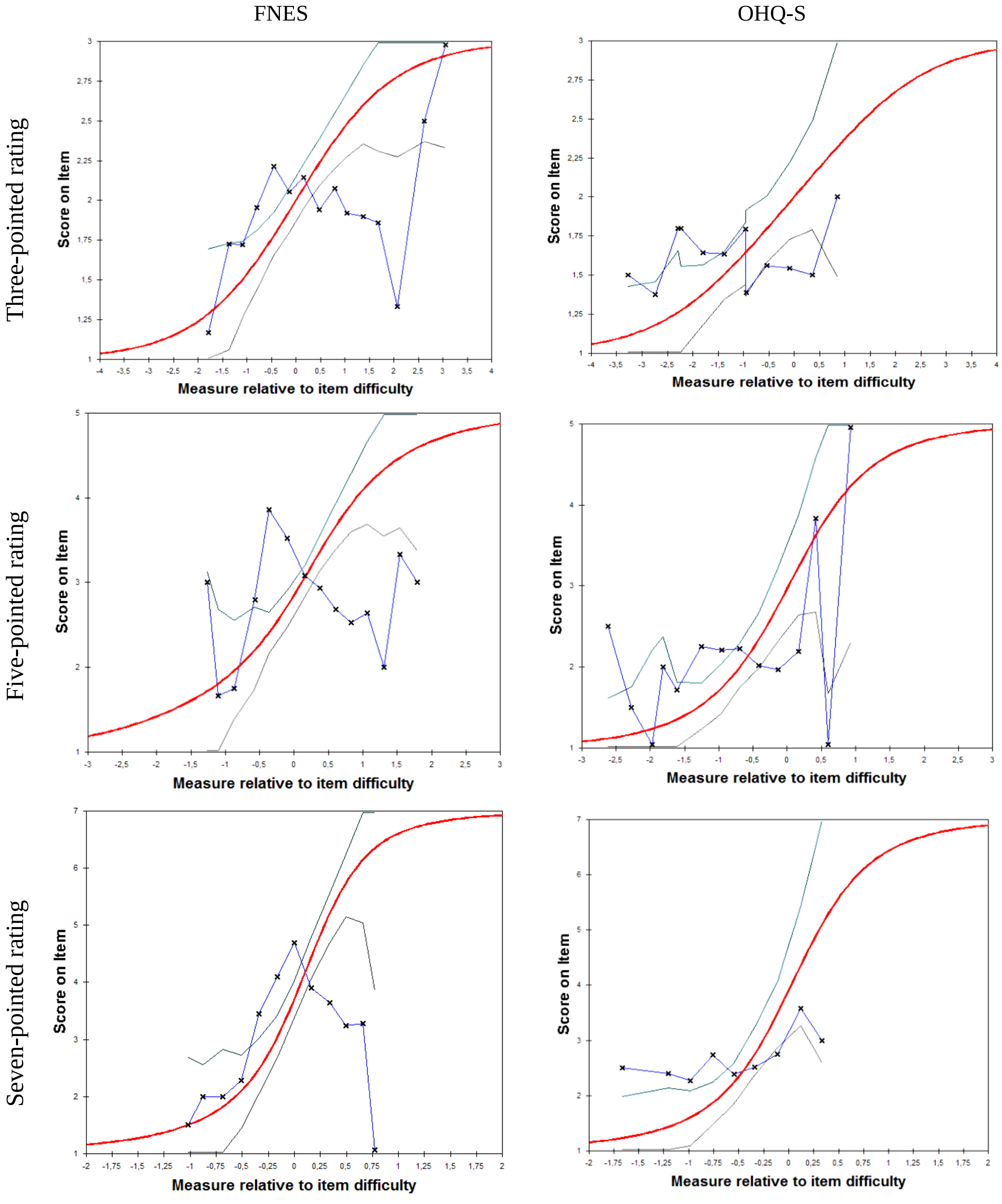


Figure 2. Test characteristic curves for reverse scored items in FNES and OHQ-S

Since category statistics reveal that the scale categories in reverse scored items cannot be distinguished by participants and since reverse scored items are found not to work in the same way as straightforward items according to test characteristic curves, the reverse scored items in FNES and OHQ-S were excluded while analysing the effects of rating type measurements on psychometric properties. Thus, reverse scored items were removed from the scale and Rasch analysis was repeated with straightforward items. Taking the confusing effects – which could stem from the fact that reverse scored items had psychometric properties different from straightforward items- under control was targeted in researching the effects of the number of response categories on validity and reliability with this study. Table 4 shows the measurement report for the facet of persons in FNES and OHQ-S in three, five and seven-pointed rating.

Table 4. The Measurement Report for the Facet of Person in FNES and OHQ-S in Three, Five and Seven-Pointed Rating

			FNES			OHQ-S		
			Measure	Infit	Outfit	Measure	Infit	Outfit
Scaling	Three	Mean	-1.58	1.02	.99	.80	.98	.94
		Standard Deviation	1.75	.58	.64	1.83	.79	.77
		Separation Ratio	1.57			1.53		
		Reliability	.71			.70		
		Chi-square (χ^2)	628.4			328.8		
		Degrees of Freedom	196			114		
	Five	Mean	-1.12	1.04	1.02	.63	1.00	.95
		Standard Deviation	1.68	.78	.78	1.38	.91	.82
		Separation Ratio	2.17			1.79		
		Reliability	.82			.76		
		Chi-square (χ^2)	897.5			359.6		
		Degrees of Freedom	196			114		
	Seven	Mean	-.94	1.05	1.02	.27	1.01	.98
		Standard Deviation	1.28	.91	.92	.85	.90	.84
		Separation Ratio	1.56			1.51		
		Reliability	.71			.70		
		Chi-square (χ^2)	743.9			285.5		
		Degrees of Freedom	196			114		

According to Table 4, there are no significant differences in the infit and outfit statistics calculated for the facet of person in FNES and OHQ-S. In all three types of rating, the infit and outfit statistics calculated for the facet of person are in the interval of .5 and 1.5- which is acceptable (Linacre, 2014). Accordingly, model-data fit can be said to be attained. Linacre (2014) states that the fit between a model and its data informs us of the validity of the data. Therefore, it may be stated that there are no significant differences between three, five and seven-pointed rating types in terms of the validity of data, and that the model-data fit is attained no matter what number of response categories is used.

An examination of separation ratios and reliability values in Table 4 shows that the values reported for three and seven-point rating types are very close. It was also found accordingly that the separation ratio for five-pointed rating and reliability values were higher than those calculated in three and seven-pointed rating. This separation ratio found for the facet of person in FNES and OHQ-S, reliability and Chi-square values show that the latent property to be measured is discriminated more successfully in five-pointed rating than in three or seven-pointed rating. After measurements for the facet of person, the measurement reports for the facet of items were analysed. The measurement report for the facet of item in FNES and OHQ-S in three, five and seven-pointed rating types are shown in Table 5.

Table 5. The Measurement Report for the Facet of Item in FNES and OHQ-S in Three, Five and Seven-pointed Rating Types

			FNES				OHQ-S			
			Measure	Infit	Outfit	Corr. PtBis	Measure	Infit	Outfit	Corr. PtBis
Scaling	Three	Mean	.00	1.00	.99	.53	.00	.99	.94	.44
		Standard Deviation	.62	.10	.15	.05	.98	.14	.15	.05
		Separation Ratio	4.13				4.83			
		Reliability	.94				.96			
		Chi-square (χ^2)	119.8				89.5			
		Degrees of Freedom	7				4			
	Five	Mean	.00	.99	1.02	.64	.00	.99	.95	.52
		Standard Deviation	.42	.22	.25	.06	.70	.24	.26	.07
		Separation Ratio	4.60				6.04			
		Reliability	.95				.97			
		Chi-square (χ^2)	149.8				119.0			
		Degrees of Freedom	7				4			
	Seven	Mean	.00	1.02	1.02	.62	.00	1.01	.98	.45
		Standard Deviation	.21	.13	.21	.04	.39	.10	.12	.06
		Separation Ratio	3.54				5.55			
		Reliability	.93				.97			
		Chi-square (χ^2)	96.0				107.1			
		Degrees of Freedom	7				4			

On examining Table 5, it is observed that the infit and outfit statistics calculated for the facet of item in FNES and OHQ-S are very close. In all three types of rating, the infit and outfit statistics are within the interval of .5 and 1.5 – which is recommended to be considered (Linacre, 2014). These values for fit statistics indicate that the model fits the data, and that the validity of the data is attained.

According to Table 5, the point biserial correlation values are available on the right of the columns of the infit and outfit statistics. These correlations are the counterpart for Pearson’s correlations (Linacre, 2014), and are considered as evidence for item discrimination (item validity). Point biserial coefficients are presented separately for each item and are also reported as an average coefficient for the overall scale in Rasch analysis outputs. However, the point biserial coefficients are not presented separately for each item in Table 5. They are shown as average values corresponding to the division of total correlation coefficients to the number of items in the scale. According to these average values, it was found that correlation coefficients for the five and seven-pointed rating in FNES were close. Almost no differences were found between point biserial correlation coefficients calculated for three and seven-pointed rating in OHQ-S. It was also found that the point biserial correlation coefficients calculated for five-pointed rating was higher than those calculated for three and seven-pointed rating. On considering the biserial correlation coefficients calculated in FNES and OHQ-S altogether, it is found that five-pointed rating yields higher correlation coefficients than three-pointed and seven-pointed rating. Therefore, it can be said that item discrimination rises when five-pointed rating is used instead of three or seven-pointed rating in Likert type scales.

On checking the reliability shown in Table 5, it is found that coefficients calculated for the three, five and seven-pointed types of rating in FNES and OHQ-S are almost the same. In other words, the number of response categories has no significant effects on item reliability. However, the separation ratio for the facet of item and the chi-square values differ according to the number of response categories. The highest values for the separation ratio and for the chi-square test in both FNES and PHQ-S were obtained in five-pointed rating. On comparing the separation ratio and chi-square test results for three-pointed and seven-pointed rating, it was found that the values for three-pointed rating were higher in FNES and that the values for seven-pointed rating were higher in OHQ-S.

Accordingly, although it looks impossible to make a clear inference as to in which type (three-pointed or seven-pointed) of rating items are discriminated better, it can be said that the items with differing levels of difficulty (items in which there are differences between the probabilities of participants' agreement and disagreement) are discriminated better in five-pointed rating.

DISCUSSION AND CONCLUSIONS

This study investigated reverse scored items and the number of response categories problem in Likert type scales. It was found accordingly that three, five and seven-pointed rating types all worked actively in straightforward items. Yet, it was also found that scale categories could not be distinguished by participants in reverse scored items no matter what number of categories was used. Not fulfilling the assumption of the *scale categories for reverse scored items were symmetrical and equi-distant* under the real conditions could be a cause for this problem (Locker, Jokovic & Allison, 2013). This finding of the research is supported by DeVellis' (2003) explanation that including reverse scored items in Likert type scales can present certain disadvantages. In the book entitled *Scale Development: Theory and Applications*, DeVellis (2003) states that participants can have confusion about what the responses *strongly agree* or *strongly disagree* mean while answering the reverse scored items in Likert Type scales. Such confusion can lead to not being able to distinguishing between scale categories in reverse scored items. Therefore, the findings of this research are aligned with the explanations made by DeVellis (2003). The findings obtained by Conrad et al (2004) are also similar to the ones obtained in this current study. In their study by Conrad et al (2004) by using Mississippi Scale for Posttraumatic Stress Disorder, they found that four out of six items violating model-data fit were reverse scored items and authors pointed out that the category statistics for those four items were problematic.

The results obtained by Bergstrom and Lunz (1998), however, differ from the ones obtained in this study. Bergstrom and Lunz (1998) administered the Job Satisfaction Scale of 36 items (19 straightforward and 17 reverse scored) to a study group containing 706 participants, and analysed the category statistics for the straightforward and reverse scored items in the scale. They found in consequence that the scale categories worked actively in both straightforward and reverse scored items. The inconsonance between results in Bergstrom and Lunz (1998) and in this study can be attributed to the different procedures followed in the two studies. The straightforward and reverse scored items in the data collection tools were analysed together in this study, and syntax was prepared on the basis of examples Linacre (2014) gave in the users' manual for FACETS programme. The numbers for the straightforward and reverse scored items in the scale, and a command to reverse score the scale categories in the reverse items were added to the syntax. In Bergstrom and Lunz (1998), on the other hand, straightforward and reverse scored items were included in two different data sets, and the analyses were performed separately for the two sets. That is to say, Bergstrom and Lunz (1998) did not analyse straightforward and reverse scored items as the items measuring the same structure, but they analysed the items as if they belonged to two different scales. This situation can be seen as the cause for the non-overlap between the results obtained in Begrstrom and Lunz (1998) and in this study.

This study found that the observed and the expected test characteristic curves in straightforward items overlapped to a large extent, but that there were significant deviations between the observed and the expected test characteristic curves in reverse scored items. Accordingly, it could be said that model-data fit was attained in straightforward items but that it was not attained in reverse scored items. This research finding showing that straightforward items served to measuring the intended latent structure but that reverse scored items failed to measure the same latent structure is supported by the findings obtained by Schriesheim and Eisenbach (1995), Meloni and Gana (2001), Conrad et al. (2004), Roszkowski and Soven (2010), Hooper et al. (2013), and Locker, Jokovic and Allison (2013). Schriesheim and Eisenbach (1995) found that the variance stemming from the property to be measured was higher in items containing positive statements than in items containing negative statements. In the study by Meloni and Gana (2001), the validity and reliability of the Italian version of Penn State Worry Questionnaire. In the study, high correlations were found between scores received from the overall scale and the straightforward items in the scale. On the other hand, the study found that the correlations between reverse scored items and the scores received from the overall scale were lower. Besides, it was also found in the above mentioned study that the correlations between scale items and the sores received from the scale of self-actualization - which was used for criterion validity in the study- were not significant in any of the reverse scored items. Meloni and Gana (2001) stated based on this finding that reverse scored items reduced the validity of a scale and that the psychometric properties of a scale would be improved by removing those items from the scale. Conrad et al (2004) analysed the limitations of reverse scored items through the Mississippi Scale for Posttraumatic Stress Disorder containing 35 items 25 of which were straightforward and 10 of which were reverse scored items. Accordingly, the study found that reverse scored items caused a reduction in model-data fit and that validity improved without loss of reliability when those items were removed from the scale. Roszkowski and Soven (2010) reported that reverse scored items had item total correlations lower than straightforward items, that those items constituted a separate factor in themselves and that Cronbach Alpha internal consistency coefficient rose on removing them. The same study also found that a distinctive increase occurred in item total correlations and a one-factor structure was obtained when reverse scored items were expressed as straightforward. Hooper et al (2013) found that reverse scored items caused reduction in model-data fit, they made a measurement model complex, and that they caused inclusion of variance irrelevant to the measured structure in measurement results. In a similar vein, Locker, Jokovic and Allison (2013) concluded that even though they were written to measure the same structure straightforward and reverse scored items were in differing factors, that items expressed in straightforward way had mutual correlations higher than those expressed in reverse scored way, and that there were significant differences between average scores received from straightforward items and the average scores received from reverse scored items. All of the studies mentioned above agree with the finding that including reverse scored items in Likert type scales would influence measurement results in negative ways. Demonstrating in many studies analysing the factor structures of Likert type scales composed of straightforward and reverse scored items (Benson & Hocevar 1985; Bolin & Dodder. 1990; Herche & Engelland 1996; Kelloway, Catano & Southwell 1992; Lai, 1994; McInerney, McInerney & Roche, 1994; Pilotte & Gable 1990; Rodebaugh et al., 2004; Spector, van Katwyk, Brannick & Chen, 1997) that reverse scored items constitute a separate factor in themselves also support the findings obtained in this research because –as Ibrahimoğlu (2001) states- gathering straightforward and reverse scored items under different factors in a scale could mean that reverse scored items cause the inclusion of variables other than the property to be measured in measurement results (cited in Weems, Onwuegbuzie & Lustig, 2003).

On examining the effects of the number of response categories used in a scale on item correlations, it was found that item discrimination was higher in five-pointed rating than in three and seven-pointed rating. This finding overlaps with the theoretical knowledge included in the article by Jacoby and Matell (1986) entitled "*Are scales with three-pointed rating good enough?"* Jacoby and Matell (1971) pointed out that a scale could not give detailed information if the number of response categories in a scale is too small and that the discrimination would decrease due to this. In their opinion, if the number of response categories is too big, on the other hand, reductions in item discrimination can occur due to the fact that participants cannot distinguish between different points of the scale. The fact that item discrimination calculated for five pointed rating in FNES and OHQ-S was higher than those calculated for three and seven-pointed rating is compatible with explanations made by Tezbaşaran (1997). In the book entitled *A Guide for Developing Likert Type Scales*, Tezbaşaran (1997) pointed out that three, five or seven-pointed rating could be used in Likert type scales, but that the most appropriate number of response categories was five.

Similar results were yielded in the three, five and seven-pointed forms of rating in FNES and OHQ-S in this study in terms of model-data fit. That is to say, it was found that the number of response categories had no significant effects on model-data fit. This finding is parallel to the one obtained by Daher, Ahmad, Winn and Selamat (2015). Daher et al (2015) analysed the data collected with three, four and six-pointed rating of Malay spiritual well-being scale according to the Rasch Model. In consequence, they found that the fit statistics calculated for all three rating types were similar and that the number of response categories did not have significant effects on model-data fit. Model-data fit is regarded as evidence for the validity of measurements in the Rasch analysis (Linacre, 2014). Therefore, the finding that the number of response categories had no considerable effects on model-data fit can be interpreted that valid measurements can be performed by using any of three, five and seven-pointed rating in a scale of items reflecting the structure to be measured.

It was found through Rasch analysis that the reported reliability for the facet of person separation ratio and Chi square rose on raising the number of response categories in the scale from three to five. This finding indicates that individuals at different levels of the latent structure to be measured are discriminated more effectively in five-pointed rating than in three-pointed rating. One of the basic factors determining how well individuals are discriminated in consequence of a measurement is the extent to which a scale is precise. As the number of response categories decreases, the sensitivity of a scale falls (Erkuş, 2012), and this fall in sensitivity can lead to a fall in reliability. Here, discrimination of individuals more effectively in five-pointed rating than in three-pointed rating can be explained with the fact that the sensitivity of measurements obtained from three-pointed rating is higher than that obtained from three-pointed rating. The study conducted by Ray (1980), which concludes that discrimination increases by raising the number of response categories from three to five, is also supportive of our findings.

The decrease in reliability values for the facet of person instead of increase when the number of response categories is raised from five to seven according to the findings reported in Rasch analysis can stem from participants' encountering problems in distinguishing between the categories in seven-pointed rating because the increase in the number of response categories in the scale can only increase sensitivity up to a certain point. And increasing the number of categories too much causes a fall in the perception of discrimination between categories (Erkuş, 2012), and as a result, this can influence reliability for the facet of person. At this point, the question of whether the number of response categories in seven-pointed rating is more than that human mind can distinguish between comes into mind. Büyüköztürk (2005) states that whether or not individuals can make discrimination carefully enough while responding to a scale of seven-pointed rating is a matter of discussion. Miller (1956), on the other hand, claims that human mind has the capacity to distinguish between seven different categories (Cited in Preston & Colman, 2000). The fact that the category statistics obtained in this study for seven-pointed rating met the assumptions necessary to say that the scale categories worked properly overlaps with Miller's (1956) claim. Accordingly, it can be said the number of response categories in seven-pointed rating is within the limits that human mind can distinguish between. In addition, since five-pointed rating is used more frequently than seven-pointed rating in Likert type scales (Lozano, García-Cueto & Muñiz, 2008), individuals can be more familiar with five-pointed rating and can discriminate between the differences in scale categories in five-pointed rating more effectively than in seven-pointed rating. This situation is thought to be the cause for higher reliability, separation ratio and Chi-square calculated for the facet of person in five-pointed rating than in seven-pointed rating.

It was found in this study that the reliability coefficients calculated for the facet of item in three, five and seven-pointed rating was almost equal. Reliability coefficients calculated for the facet of item in the Rasch analysis correspond to Cronbach Alpha internal consistency coefficients calculated in the CTT (Linacre, 2014). Therefore, it may be said that the studies demonstrating that the number of response categories have no significant effects on Cronbach Alpha internal consistency coefficients (Aiken, 1983; Leung, 2011; Matell & Jacoby, 1971; Preston & Colman, 2000; Qasem, Almoshigah & Gupta, 2014; Wong, Peng, Shi & Mao, 2011) are all supportive of our findings obtained in this study. In contrast to the above listed studies, there are also studies conflicting with those findings in the literature. The studies conducted by Weng (2004), Lozano, García-Cueto and Muñiz (2008), Maydeu-Olivares et al. (2009), Uyumaz (2013) and Tarka (2015) and reporting that Cronbach Alpha internal consistency coefficients rise as the number of response categories in a scale increases differ from this study in terms of their findings. According to Fabiola, Iwin, Jennifer and Zaira (2012), the inconsistencies observed in the research findings concerning the effects of the number of response categories on validity and reliability can stem from the differences in the measurement models (CTT or Item Response Theory) used. For this reason, it can be said that this study analysing the effects of the number of response categories on the psychometric properties of measurements through Rasch analysis is not adequate on its own to make clear inferences about the correlations between the number of response categories and item reliability. It is predicted that clearer statements will be made about how item reliability is affected by the number of response categories with an increase in the number of studies to be made through methods based on item response theory.

This study found that the separation ratios and Chi square calculated for the facet of items in five-pointed rating were higher than the values for three and seven-pointed rating. Based on this finding, it can be said that the items with differing levels of difficulty (the likelihood of participants' agreement or disagreement) can be discriminated better in five-pointed rating than in three or seven-pointed rating. It is believed that this result is related with the sensitivity of the rating used the scale and with how effectively scale categories are discriminated by participants- as in the separation ratio and Chi square for the facet of person. As the measurement reports and category statistics for the facet of person in FNES and OHQ-S indicate, five-pointed rating can yield more sensitive measurements than three-pointed rating, and it is composed of response categories through which participants are discriminated more easily than seven-pointed rating. This property might have made five-pointed rating more effective in discriminating between items of differing difficulty than three and five-pointed rating. Compared to three-pointed and seven-pointed rating, five-pointed rating has higher separation ratio and chi square values for the facet of item- which is a finding compatible with all finding except for item reliability obtained in this study. The fact that there were differences in separation ratios and chi square in favour of five-pointed rating despite the absence of differences in item reliability between three, five and seven-pointed rating could be attributed to the fact that reliability, separation ratio and chi square were the measurements reported in different metrics. While item reliability can take on values between 0 and 1, separation ratio can take on values ranging between 1 and ∞, and Chi square can be ranged from 0 to ∞ (Sudweeks, Reeveb & Bradshawc, 2004). Therefore, it can be more difficult for some difficulties to be manifested in reliability coefficients than in separation ratio and Chi square values.

To sum up the conclusions reached in this study, it was found that the scale categories in reverse scored items could not be discriminated by responders no matter which type of rating (three, five or seven-pointed) was used, and that reverse scored items did not measure the same latent structure as straightforward items did. Considering these results showing that reverse scored items made measurement models more complicated, preparing Likert type scales having only straightforward items can be evaluated as an application which can improve the psychometric properties of measurements. This study also found that the number of response categories did not have any effects on model-data fit. On the other hand, category statistics, item discrimination, reliability coefficients and Chi square calculated for the facets of person and items demonstrated that five-pointed rating was more functional than three or seven-pointed rating. This result leads to the recommendation that five-pointed rather than three or seven-pointed rating should be preferred. Yet, the restrictions of the study limit the generalizability of the findings and they also require that the recommendation should be interpreted in the framework of these restrictions. The restrictions of the study and the recommendations to be made for further research in accordance with the restrictions are as in the following.

LIMITATIONS AND RECOMMENDATIONS

The first restriction of this study has to do with the properties of the study group. The study was conducted with a group composed of university students. The best number of response categories for a scale can differ according to participants' age and level of education (Adelson & McCoach, 2010; Fabiola et al. 2012; Tekindal, 2009). Therefore, it may be recommended that such a study be conducted with participants of different age groups and educational levels. Also replication of a similar study on different samples from Turkey will lend the generalization of the research findings to Turkish culture. The second restriction of the study is the way the reverse scored items in the scales used in this study are revealed. Reverse scored items can be formed by using words with opposite meaning as well as using negative prefixes (or suffixes) (Sonderen, Sanderman and Coyne, 2013). According to the results Swain, Weathers and Niedrich (2008) obtained by analysing approximately 2000 scale items, reverse scored items are stated by using negative prefixes or suffixes by 81%. Therefore, this study preferred the scales having reverse scored items expressed by using negative prefixes or suffixes. Because this situation restricted the generalizability of the research findings, it could be recommended that a similar study be performed by using scales with reverse scored items which are stated in words with opposite meanings. And finally, the data for this study were collected through FNES and OHQ-S, and the number of reverse scored items in both scales is about one third of straightforward items. Using equal number of straightforward and reverse scored items or more reverse scored items in prospective studies might contribute to the generalizability of the findings.

REFERENCES

Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-type scale. *Educational and Psychological Measurement*, 70(5), 796-807. http://dx.doi.org/10.1177/0013164410366694

Ahlawat, K. S. (1985). On the negative valence items in self-report measures. *The Journal of General Psychology*, *112*(1), 89-99. http://dx.doi.org/10.1080/00221309.1985.9710992

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement*, *43*(2), 397-401. http://dx.doi.org/10.1177/001316448304300209

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *The Public Opinion Quarterly*, *48*(2), 491-509. http://dx.doi.org/10.1086/268845

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Barnette, J. J. (1999, April). *Likert response alternative direction: SA to SD or SD to SA: Does it make a difference?* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada. Retrieved from http://eric.ed.gov/?id=ED429125

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, *22*(3), 231–240. http://dx.doi.org/10.1111/j.1745-3984.1985.tb01061.x

Bergstrom, B. A., & Lunz, M. E. (1998, April). *Rating scale analysis: Gauging the impact of positively and negatively worded items*. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA. Retrieved from http://files.eric.ed.gov/fulltext/ED423289.pdf

Birkett, N. J. (1986). *Selecting the number of response categories for a Likert-type scale*. Retrieved from http://www.amstat.org/sections/srms/Proceedings/papers/1986_091.pdf

Bolin, B. L., & Dodder, R. A. (1990). The affect balance scale in an American college population. *The Journal of Social Psychology*, *130*(6), 839-40. http://dx.doi.org/10.1080/00224545.1990.9924639

Büyüköztürk, Ş. (2005). Anket geliştirme. *Türk Eğitim Bilimleri Dergisi, 3*(2), 133-151. Retrieved from http://www.tebd.gazi.edu.tr/index.php/tebd/article/view/315/297

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied Psychological Measurement, 9*(1), 31-36. http://dx.doi.org/10.1177/014662168500900103

Chamberlain, V. M., & Cummings, M. N. (1984). Development of an instructor/course evaluation instrument. *College Student Journal, 18*(3), 246-250.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205-215. http://dx.doi.org/10.1177/014662169401800302

Chiorri, C., Anselmi, P., & Robusto, E. (2009). Reverse items are not opposites of straightforward items. In U. Savardi (Ed.), *The perception and cognition of contraries* (pp. 295-328). Milano: McGraw-Hill.

Comrey, A. L., & Montang, I. (1982). Comparison of factor analytic results with two choice and seven choice personality item formats. *Applied Psychological Measurement, 6*(3), 285-289. http://dx.doi.org/10.1177/014662168200600304

Conrad, K. J., Wright, B. D., McKnight, P., McFall, M., Fontana A., & Rosenheck, R. (2004). Comparing traditional and Rasch analyses of the Mississippi PTSD scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement, 5*(1), 15-30. Retrieved from https://www.academia.edu/2832927/Comparing_traditional_and_Rasch_analyses_of_the_Mississippi_PTSD_scale_Revealing_limitations_of_reverse-scored_items

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3-31. http://dx.doi.org/10.1177/001316445001000101

Çetin, B., Doğan, T. ve Sapmaz, F. (2010). Olumsuz değerlendirilme korkusu ölçeği kısa formu'nun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Eğitim ve Bilim, 35*(156), 205-216.

Daher, A. M., Ahmad, S. H., Winn, T., & Selamat, M. I. (2015). Impact of rating scale categories on reliability and fit statistics of the Malay spiritual well-being scale using Rasch analysis. *Malaysian Journal of Medical Sciences, 22*(3), 48-55. Retrieved from http://www.bioline.org.br/pdf?mj15032

Dawes, J. (2007). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*(1), 61-77. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.417.9488&rep=rep1&type=pdf

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Newbury Park: Sage.

Doğan, T. ve Akıncı Çötok, N. (2011). Oxford mutluluk ölçeği kısa formunun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 4*(36), 165-172. Retrieved from http://dergipark.ulakbim.gov.tr/tpdrd/article/view/1058000176/1058000178

Erkuş, A. (2003). *Psikometri üzerine yazılar*. Ankara: Türk Psikologlar Derneği Yazıları.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-I*. Ankara: Pegem Akademi.

Fabiola, G. B., Iwin, L., Jennifer, L. M., & Zaira, V. V. (2012). The effect of the number of answer choices on the psychometric properties of stress measurement in an ınstrument applied to children. *Evaluar, 12* 43-59. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar/article/download/4694/4488

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi- square fit indices in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 4*(2), 108-120, http://dx.doi.org/10.1080/10705519709540064

Güler, N., İlhan, M., Güneyli, A., & Demir, S. (2017). An evaluation of the psychometric properties of three different forms of Daly and Miller's writing apprehension test through Rasch analysis. *Educational Sciences: Theory & Practice, 17*(3), 721-744. http://dx.doi.org/10.12738/estp.2017.3.0051

Halpin, G., Halpin, G., & Arbet, S. (1994). Effects of number and type of response choices on internal consistency reliability. *Perceptual and Motor Skills, 79*(2), 928-930. http://dx.doi.org/10.2466/pms.1994.79.2.928

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science, 24*(4), 366-374. http://dx.doi.org/10.1177/0092070396244007

Hofstede, G. (1998). *Masculinity and femininity: The taboo dimension of national cultures*. Thousand Oaks, CA: Sage.

Hooper, M., Arora, A., Martin, M. O., & Mullis, I. V. S., (2013, June). *Examining the behavior of "reverse directional" items in the TIMSS 2011 context questionnaire scales*. Paper Presented at the 5th IEA International Research Conference. National Institute of Education, Nanyang Technological University, Singapore. Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Hooper_etal.pdf

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309. http://dx.doi.org/10.1177/0022022189203004

Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*(2), 497-500. http://dx.doi.org/10.2466/pr0.2001.88.2.497

Jacoby, J., & Matell, M. S. (1971). Three-point likert scales are good enough. *Journal of Marketing Research, 8*, 495-500. Retrieved from https://www.jstor.org/stable/pdf/3150242.pdf?_=1472027712885

Jenkins, G. D., & Taber, T. D. (1977). A Monte-Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*(4), 392-398. http://dx.doi.org/10.1037/0021-9010.62.4.392

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264-277. http://dx.doi.org/10.1177/0022022104272905

Kelloway, E. K., Catano, V. M., & Southwell, R. R. (1992). The construct validity of union commitment: Development and dimensionality of a shorter scale. *Journal of Occupational and Organizational Psychology, 65*(3), 197-211. http://dx.doi.org/10.1111/j.2044-8325.1992.tb00498.x

Kim, K. H. (1998). An analysis of optimum number of response categories for Korean consumers. *Journal of Global Academy of Marketing Science*, *1*(1), 61-86. http://dx.doi.org/10.1080/12297119.1998.9707386

King, L. A., King, D., & Klockars, A. J. (1983). Dichotomous and multipoint scales using bipolar adjectives. *Applied Psychological Measurement*, *7*(2), 173-180. http://dx.doi.org/10.1177/014662168300700205

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). New York, NY: Routledge.

Lai, J. C. L. (1994). Differential predictive power of the positively versus the negatively worded items of the life orientation test. *Psychological Repors*, *75*(3), 1507-1515. http://dx.doi.org/10.2466/pr0.1994.75.3f.1507

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health*, *25*(4), 295-306. http://dx.doi.org/10.1002/nur.10041

Leung, S. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, *37*(4), 412-421. http://dx.doi.org/10.1080/01488376.2011.580697

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 2-55.

Linacre, J. M. (2014). *A user's guide to FACETS Rasch-model computer programs.* Retrieved from http://www.winsteps.com/a/facets-manual.pdf

Lissitz, R. W., & Green, S. B. (1975). Effects of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*(1), 10-13. http://dx.doi.org/10.1037/h0076268

Locker, D., Jokovic, A., & Allison, P. (2013). Direction of wording and responses to items in oral health-related quality of life questionnaires for children and their parents. *Community Dent Oral Epidemiol 35*(4), 255-262. http://dx.doi.org/10.1111/j.1600-0528.2007.00320.x

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 4(2), 73-79. http://dx.doi.org/10.1027/1614-2241.4.2.73

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*(3), 657-674. http://dx.doi.org/10.1177/001316447103100307

Maydeu-Olivares A., Kramp U., García-Forero C., Gallardo-Pujol, D., Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, *41*(2), 295-308. http://dx.doi.org/10.3758/BRM.41.2.295

McInerney, V., McInerney, D., & Roche, L. (1994, July). Definitely not just another computer anxiety instrument: The development and validation of CALM: Computer anxiety and learning measure. Paper presented at the Annual Stress and Anxiety Research Conference, Madrid, Spain. Retrieved from http://files.eric.ed.gov/fulltext/ED386161.pdf

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*(2), 549-550. http://dx.doi.org/10.2466/pms.1989.68.2.549

Østerås, N., Gulbrandsen, P., Garratt, A., Benth, J. S., Dahl, F. A, Natvig, B., & Brage, S. (2008). A randomised comparison of a four- and a five-point scale version of the Norwegian function assessment scale. *Health and Quality of Life Outcomes*, *6*(14), 1-9, http://dx.doi.org/10.1186/1477-7525-6-14

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*(3), 603-610. http://dx.doi.org/10.1177/0013164490503016

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1-15. http://dx.doi.org/10.1016/S0001-6918(99)00050-5

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*(4), 513-533. http://dx.doi.org/10.1007/BF02291492

Ray, J. (1980). How many answer categories should attitude and personality scales use? *South African Journal of Psychology*, *10*, 53-54. Retrieved from http://jonjayray.tripod.com/howmany.html

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment*, *16*, 169-181. http://dx.doi.org/10.1037/1040-3590.16.2.169

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*(1), 113-130. http://dx.doi.org/10.1080/02602930802618344

Qasem, M., Almoshigah, T., & Gupta, S. (2014). The effect of number of alternatives on validity and reliability in Likert scale. *International journal of innovative research & studies, 3*(6), 324-333. http://dx.doi.org/10.13140/2.1.2237.2803

Schrieheim, C. A, & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114. http://dx.doi.org/10.1177/001316448104100420

Spector, P. E, van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How Item characteristics can produce artifactual factors. *Journal of Management, 23*(5), 659-677. http://dx.doi.org/10.1016/S0149-2063(97)90020-9

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *Journal of Social Psychology, 122*(2), 151-156. http://dx.doi.org/10.1080/00224545.1984.9713475

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. http://dx.doi.org/10.1016/j.asw.2004.11.001

Swain S. D, Weathers D., Niedrich R. W. (2008) Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research 45*, 116-131. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=990097

Şeker, H. ve Gençdoğan, B. (2006). *Psikolojide ve eğitimde ölçme aracı geliştirme*. Ankara: Nobel.

Tarka, P. (2015). Likert scale and change in range of response categories vs. the factors extraction in EFA model. *Folia Oeconomica, 1*(311), 27-36. http://dx.doi.org/10.18778/0208- 6018.311.04

Taşdelen Teker, G., Güler, N., & Kaya Uyanık, G. (2015). Comparing the effectiveness of SPSS and EduG using different designs for generalizability theory. *Educational Sciences: Theory & Practice, 15*(3), 635-645. http://dx.doi.org/10.12738/estp.2015.3.2278

Tavşancıl, E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Ankara: Nobel.

Tekindal, S. (2009). *Duyuşsal özelliklerin ölçülmesi için araç oluşturma*. Ankara: Pegem Akademi.

Tezbaşaran, A. (1997). *Likert tipi ölçek hazırlama kılavuzu*. Ankara: Türk Psikologlar Derneği.

Turan, İ., Şimşek, Ü. ve Aslan, H. (2015). Eğitim araştırmalarında Likert ölçeği ve Likert tipi soruların kullanımı ve analizi. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi, 30*, 186-203. Retrieved from http://dergipark.ulakbim.gov.tr/sakaefd/article/view/5000143504

Van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's Learn from cows in the rain. *PloS one, 8*(7), 1-7. http://dx.doi.org/10.1371/journal.pone.0068967

Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. (2003). Profiles of respondents who respond inconsistently to positively- and negatively- worded items on rating scales. *Evaluation & Research in Education, 17*(1), 45-60. http://dx.doi.org/10.1080/14664200308668290

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972. http://dx.doi.org/10.1177/0013164404268674

Wong, C. S., Peng, K. Z., Shi J., & Mao, Y. (2011). Differences between odd number and even number response formats: Evidence from mainland Chinese respondents. *Asia Pacific Journal of Management, 28*(2), 379–399. http://dx.doi.org/10.1007/s10490-009-9143-6

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point likert type response scales. *Educational and Psychological Measurement, 47*(1), 27-35. http://dx.doi.org/10.1177/0013164487471003

Zhang, X., Noor, R., & Savalei, V. (2016) Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PLoS ONE, 11*(6), 1-15. http://dx.doi.org/10.1371/journal.pone.0157795

UZUN ÖZET

Giriş

Bu çalışmanın iki temel amacı bulunmaktadır: Bunlardan ilki; Likert tipi ölçeklerdeki olumsuz maddelerin ne derece işlevsel olduğunun tespit edilmesidir. Bu amaç doğrultusunda araştırmada; *i)* olumlu ve olumsuz maddelerde ölçek kategorilerinin aynı şekilde çalışıp çalışmadığı incelenmiş, *ii)* olumlu ve olumsuz maddelere ait test karakteristik eğrileri karşılaştırılarak bu maddelerin aynı örtük yapıyı ölçüp ölçmedikleri belirlenmeye çalışılmıştır. Belirtilen işlemler üç, beş ve yedili derecelemeye sahip Likert tipi ölçekler için ayrı ayrı gerçekleştirilmiştir. Böylelikle olumsuz maddelerin işleyişinin ölçekteki kategori sayısından etkilenip etkilenmediği kontrol edilmiştir. Araştırmanın ikinci temel amacını Likert tipi ölçeklerde kullanılan kategori sayısının ölçümlerin psikometrik özellikleri üzerindeki etkisinin ortaya konulması oluşturmaktadır. Bu doğrultuda; üç, beş ve yedili derecelendirmeye sahip Likert tipi ölçekler güvenirlik ile model-veri uyumu açısından karşılaştırılmıştır. Bu sayede kategori sayısının güvenirlik üzerindeki etkisi incelenirken yapı geçerliği de göz ardı edilmemiştir. Bunun araştırma sonuçlarını anlamlı kılma adına oldukça önemli olduğu düşünülmektedir. Çünkü Cronbach'ın (1950) da belirttiği gibi yalnızca güvenirliği arttırmanın tek başına bir değeri bulunmamakta; güvenirliği arttıran bir kategori sayısının uygun olduğunun söylenebilmesi için geçerliğin de dikkate alınması gerekmektedir.

Yöntem

Araştırma, toplamda 312 üniversite öğrencisinden oluşan iki ayrı çalışma grubu üzerinde yürütülmüştür. Birinci çalışma grubunda 197 ve ikinci çalışma grubunda 115 katılımcı yer almıştır. Çalışmada veri toplama aracı olarak Olumsuz Değerlendirilme Korkusu Ölçeği (ODKÖ) ile Oxford Mutluluk Ölçeği-Kısa Formu (OMÖ-K) kullanılmıştır. Leary (1983) tarafından geliştirilip; Çetin, Doğan ve Sapmaz (2010) tarafından Türkçeye uyarlanan ODKÖ sekizi olumlu (olumsuz değerlendirilme korkusunu destekleyen) ve üçü olumsuz (olumsuz değerlendirilme korkusunu desteklemeyen) toplam 11 madde içermektedir. OMÖ-K ise Hills ve Argyle (2002) tarafından geliştirilmiş, Doğan ve Akıncı Çötok (2011) tarafından Türkçe'ye uyarlanmıştır. Bu ölçekte beşi olumlu ve ikisi olumsuz toplam yedi madde bulunmaktadır. Birinci çalışma grubundaki katılımcılara ODKÖ, ikinci çalışma grubundaki katılımcılara OMÖ-K üç, beş ve yedili dereceleme ile uygulanmıştır. Her üç derecelemede de kategorilerin yalnızca uç noktaları isimlendirilmiş (*Hiç Katılmıyorum → Tamamen Katılıyorum*); uç noktalar arasında kalan seçenekler için bir adlandırma kullanılmamıştır. Bu tür bir yaklaşımın benimsenmesinde, yedili derecelemede, üçlü ve beşli derecelemedeki kadar net bir isimlendirme yapılamayacağı düşüncesi etkili olmuştur. Şöyle ki, Likert tipi ölçeklerde ölçek noktalarının ne kadar net bir biçimde adlandırıldığına bağlı olarak ölçme sonuçlarında farklılıklar gözlenebilmektedir. Bu bakımdan üçlü, beşli ve yedili derecelemede tüm kategoriler için isimlendirme kullanılması halinde araştırma sonucunda ulaşılan bulguların gerçekten kategori sayısındaki farklılıktan mı; yoksa kategorilere ilişkin adlandırmaların aynı kesinlikte olmayışından mı kaynaklandığını belirlemek mümkün olmayacaktır. Bu noktadan hareketle çalışmada; her üç ölçek formunda da (hem üç, hem beş hem de yedili derecelemede) kategorilerin sadece uç noktaları isimlendirilmiştir. Araştırma kapsamında toplanan veriler FACETS paket programından yararlanılarak Rasch modeline göre analiz edilmiştir.

Sonuç ve Tartışma

Rasch analizinden elde edilen bulgular, ODKÖ ile OMÖ-K'deki düz puanlanan maddelerde gözlenen ve beklenen test karakteristik eğrilerinin büyük ölçüde örtüştüğünü, her üç dereceleme türünün de etkin bir biçimde çalıştığını ve ölçek kategorileri arasındaki farkların katılımcılar tarafından başarılı bir biçimde ayırt edildiğini ortaya koymuştur. Diğer taraftan ters puanlanan maddelerde gözlenen ile beklenen test karakteristik eğrileri arasında önemli farklılıklar olduğu ve ölçek kategorilerinin etkin bir biçimde çalışmadığı saptanmıştır. Üç, beş ve yedili derecelendirmeden hangisi kullanılırsa kullanılsın katılımcıların ters puanlanan maddelerde ölçek kategorilerini birbirinden ayırt edemediği belirlenmiştir. Olumsuz maddelerin ölçme modelini karmaşıklaştırdığı gösteren bu sonuçlar dikkate alındığında Likert tipi ölçeklerin sadece olumlu maddeleri içerecek şekilde hazırlanması, ölçümlerin psikometrik özelliklerinin iyileşmesine katkı sağlayacak bir uygulama olarak değerlendirilebilir. Araştırmada ayrıca, kategori sayısının model-veri uyumu üzerinde önemli bir etkisinin olmadığı tespit edilmiştir. Madde ayırt ediciliği, birey yüzeyine ilişkin güvenirlik katsayısı ile birey ve madde yüzeyleri için hesaplanan ayırma oranı ve Ki Kare değerlerinin ise beşli derecelemede üçlü ve yedili derecelemeye kıyasla daha yüksek olduğu sonucuna ulaşılmıştır. Bu sonuç, Likert tipi ölçeklerde üçlü veya yedili derecelemedense beşli derecelemenin tercih edilmesi şeklinde bir öneriyi beraberinde getirmektedir. Ancak, araştırmanın sınırlılıkları çalışmadan elde edilen bulguların genellenebilirliğini kısıtladığı gibi getirilen önerilerin de bu sınırlılıklar çerçevesinde yorumlanmasını gerekli kılmaktadır. Çalışmaya ilişkin sınırlılıklar ve bu sınırlılıklar doğrultusunda getirilebilecek ileri araştırma önerileri şu şekilde sıralanabilir.

Araştırmanın sınırlılıklarından ilki, çalışma grubunun özellikleri ile ilgilidir. Araştırma üniversite öğrencilerinden oluşan bir çalışma grubu üzerinde yürütülmüştür. Ölçek için en uygun kategori sayısı, katılımcıların yaşı ve eğitim düzeyine göre farklılık gösterebilmektedir. Dolayısıyla, bu tür bir çalışmanın farklı yaş gruplarından ve eğitim seviyelerinden katılımcılarla yapılması önerilebilir. Ayrıca, benzer bir çalışmanın Türkiye'den farklı örneklemler üzerinde tekrarlanması araştırma bulgularının Türk kültürüne genellenebilirliğini arttırması bakımından önem taşımaktadır. Çalışmada kullanılan ölçeklerdeki olumsuz maddelerin ifade ediliş şekilleri, araştırmaya ilişkin ikinci bir sınırlılıktır. Olumsuz maddeler, olumsuzluk ekleri (-me, -ma ve değil gibi) kullanılarak yazılabildiği gibi zıt anlamlı kelimeler kullanılarak da oluşturulabilmektedir. Bu çalışmada olumsuzluk ekleriyle ifade edilmiş olumsuz maddelerin yer aldığı ölçekler kullanılmıştır. Bu durum, olumsuz maddelerle ilgili araştırmada ulaşılan bulguların genellenebilirliğini kısıtladığından benzer bir çalışmanın zıt anlamlı kelimelerle oluşturulan olumsuz maddelerin yer aldığı ölçekler kullanılarak gerçekleştirilmesi önerilebilir. Son olarak bu araştırmanın verileri ODKÖ ve OMÖ ile toplanmış olup bu ölçeklerin her ikisinde de olumsuz madde sayısı olumlu madde sayısının yaklaşık üçte biri kadardır. Konu ile ilgili yapılacak ileri araştırmalarda olumlu ve olumsuz madde sayısının eşit ya da olumsuz maddelerin sayıca olumlu maddelerden fazla olduğu ölçeklerin kullanılması, çalışmadan ulaşılan bulguların genellenebilirliğine katkı sağlayabilir.